**НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ**
**«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ**
**імені ІГОРЯ СІКОРСЬКОГО»**

**Інститут прикладного системного аналізу**

**Кафедра математичних методів системного аналізу**

До захисту допущено:

В.о.завідувача кафедри

_____ Оксана ТИМОЩУК

«___»_____20__ р.

# Дипломна робота

**на здобуття ступеня бакалавра**

**за освітньо-професійною програмою «Системний аналіз і управління»**

**спеціальності 124 «Системний аналіз»**

**на тему: «Система прийняття рішень в кредитуванні на основі методів машинного навчання»**

Виконав:
студент IV курсу, групи КА-64
Печериця Віктор Олександрович _____

Керівник:
професор, д. т. н., професор кафедри ММСА
Данилов Валерій Якович _____

Консультант з основного розділу:
професор, д. т. н., професор кафедри інформаційної безпеки ФТІ
Качинський Анатолій Броніславович _____

Консультант з економічного розділу:
доцент, к.е.н., доцент кафедри ТТПЕ
Шевчук Олена Анатоліївна _____

Консультант з нормоконтролю:
доцент, к.т.н., доцент кафедри ММСА
Коваленко Анатолій Єпіфанович _____

Рецензент:
доц., к. т. н., с.н.с.
Кисельов Геннадій Дмитрович _____

Засвідчую, що у цій дипломній роботі
немає запозичень з праць інших авторів
без відповідних посилань.
Студент _____

Київ – 2020 року

**Національний технічний університет України**

**«Київський політехнічний інститут імені Ігоря Сікорського»**

**Інститут прикладного системного аналізу**

**Кафедра математичних методів системного аналізу**

Рівень вищої освіти – перший (бакалаврський)

Спеціальність – 124 "Системний аналіз"

Освітньо-професійна програма «Системний аналіз і управління»

ЗАТВЕРДЖУЮ
В. о. завідувача кафедри
_____ О.Л. Тимощук
«___»_____20__ р.

**ЗАВДАННЯ**

**на дипломну роботу студенту**

**Печериці Віктора Олександровича**

1. Тема роботи «Система прийняття рішень в кредитуванні на основі методів машинного навчання», керівник роботи Данилов Валерій Якович, д.т.н., професор, затверджені наказом по університету від «___»_____ 20__ р. №_____

2. Термін подання студентом роботи _____

3. Вихідні дані до роботи

Дані кредитної компанії «LendingClub» с 2007 по 2020 рік.

4. Зміст роботи

Побудова системи прийняття рішень для прогнозування результату кредиту на основі методів машинного навчання.

5. Перелік ілюстративного матеріалу (із зазначенням плакатів, презентацій тощо) _____

6. Консультанти розділів роботи

| Розділ | Прізвище, ініціали та посада консультанта | Підпис, дата | |
|---|---|---|---|
| | | завдання видав | завдання прийняв |
| Економічний | Шевчук О.А., доцент | | |
| Основний | Качинський А. Б. професор | | |

7. Дата видачі завдання _____

Календарний план

| № з/п | Назва етапів виконання дипломної роботи | Термін виконання етапів роботи | Примітка |
|---|---|---|---|
| 1 | Формулювання тематики (напряму) дослідження. | 03.09.2019 – 30.09.2019 | |
| 2 | Аналіз актуальності задач стосовно тематики дослідження | 01.10.2019 – 30.10.2019 | |
| 3 | Аналіз відомих результатів стосовно тематики дослідження | 01.11.2019 – 30.11.2019 | |
| 4 | Формулювання задач дослідження | 01.12.2019 – 30.12.2019 | |
| 5 | Уточнення теми дипломної роботи | 25.02.2019 | |
| 6 | Збір статичних даних, попередній аналіз даних | 01.03.2020 – 30.03.2020 | |
| 7 | Розробка програмного продукту для виконання обчислювальних експериментів | 01.03.2020 – 30.04.2020 | |
| 8 | Виконання обчислювальних експериментів, аналіз та оформлення результатів | 01.05.2020 – 20.05.2020 | |
| 9 | Оформлення пояснювальної записки у цілому | 21.05.2020 – 31.05.2020 | |
| 10 | Підготовка презентації для захисту | 28.05.2020 – 01.06.2020 | |
| 11 | Попередній захист дипломної роботи | 01.06.2020 – 03.06.2020 | |
| 12 | Захист дипломної роботи | 15.06.2020 – 18.06.2020 | |

Студент Віктор ПЕЧЕРИЦЯ
Керівник Валерій ДАНИЛОВ

# РЕФЕРАТ

Дипломна робота: 81 с., 46 рис., 10 табл., 2 додатки, 4 джерела.

Об’єкт дослідження – кредитні випадки, представлені статистичними даними.

Предмет дослідження – процес аналітичного прогнозування, що полягає у реалізації методів обробки статистичних даних і моделей машинного навчання на основі цих статистичних даних, а також аналізу результатів.

Мета роботи – побудова системи прийняття рішень для прогнозування результату кредиту на основі методів машинного навчання.

У роботі розглядається процес аналітичного прогнозування результату кредитів на основі методів машинного навчання. Особлива увага приділяється підготовці даних і створенні таких моделей, як дерева рішень і наївна модель Байєса. Наведено огляд сучасних моделей машинного навчання і методику їх побудови, процес підготовки даних і огляд даних щодо кредитних випадків і побудова моделей на основі цих даних. Запропоновано найефективнішу модель машинного навчання для побудови на її основі системи прийняття рішень.

КРЕДИТУВАННЯ, АНАЛІТИЧНЕ ПРОГНОЗУВАННЯ КРЕДИТНИХ ВИПАДКІВ, МОДЕЛІ МАШИННОГО НАВЧАННЯ, АНАЛІЗ ЯКОСТІ МОДЕЛЕЙ.

# ABSTRACT

Bachelor’s thesis: 81 p., 46 fig., 10 tabl., 2 appendixes, 4 sources.

The object of research – loan cases, presented by statistical data.

The subject of research – methods of data preparation and machine learning algorithms for loan result prediction.

The purpose of the research – creating a decision making system based on machine learning algorithms for loan result prediction.

The paper considers the process of analytical prediction of loan results based on machine learning methods. Particular attention is paid to data preparation and the creation of models such as decision trees and the naive Bayesian model. An overview of modern models of machine learning and methods of their construction, the process of data preparation and review of data on credit cases and the construction of models based on these data. The most effective model of machine learning for building a decision-making system based on it is proposed.

CREDITING, ANALYTICAL FORECASTING OF CREDIT CASES, MACHINE LEARNING MODELS, ANALYSIS OF MODEL QUALITY.

ЗМІСТ

# ВСТУП

Сучасні організації збирають величезні об’єми даних. Для того, щоб дані були корисними для організації, їх необхідно аналізувати та робити висновки з них, які можна використовувати для прийняття рішень. Витяг інформації з даних — задача аналізу даних. У цій роботі увага зосереджена на аналітичному прогнозуванні, яке є важливим різновидом аналізу даних.

Аналітичне прогнозування — це побудова та використання прогностичних моделей, отриманих із даних. Модель використовується для прогнозування, щоб допомогти людині і організації прийняти рішення. Ми будемо використовувати термін прогноз. У повсякденному житті ми розуміємо прогноз як те, що може статися або не статися в майбутньому. В аналітичному прогнозуванні під прогнозом слід розуміти присвоєне значення будь якої невідомої змінної. Наприклад, прогноз ціни товару, оцінка ризику, класифікація ситуації тощо. Прогностична модель навчається, щоб робити прогнози на основі набору статистичних даних. Для навчання цих моделей ми використовуємо машинне навчання.

Машинне навчання — це автоматизований процес, котрий витягає шаблони із даних. Для побудови моделей в аналітичному прогнозуванні ми використовуємо машинне навчання з учителем. Ці методи автоматично будують модель взаємозв’язків між набором описових ознак та цільовою ознакою на основі набору прецедентів – зафіксованих випадків. Після цього ми можемо використовувати цю модель для прогнозування.

Використання моделей машинного навчання є поширеною практикою в процесі кредитування, адже точність прогнозів напряму впливає на дохід організації.

# РОЗДІЛ 1. ДАНІ ПРО КРЕДИТНІ ВИПАДКИ

Процес аналітичного прогнозування складається з трьох основних частин:

1. Підготовка даних;
2. Моделювання;
3. Аналіз результатів та їх візуалізація.

Варто зазначити, що аналіз даних в організації також може включати в себе збір даних, презентування результатів, розробку програмного забезпечення в залежності від поставлених задач. В цьому розділі увага зосереджена на першому етапі аналітичного прогнозування – підготовці даних.

Перед початком роботи проведемо короткий огляд організації і даних з якими будемо працювати.

Компанія «LendingClub» - американська кредитна компанія із штаб-квартирою в Сан-Франциско, штат Каліфорнія, працює з 2006 року. Компанія заявляє, що до 2016 року вона видала кредитів на 16 мільярдів доларів. «LendingClub» дозволяє клієнтам взяти незабезпечений кредит (кредит без застави) від 1000 до 40 000 доларів, стандартний термін кредиту – три роки.

Отримані дані містять повну історію кредитів компанії за 2007-2020 роки, включаючи поточний статус кредиту і останню інформацію платежів. Інформація про компанію і її дані отримані з [4].

## 1.1 Аналітичний цикл обробки даних щодо кредитних випадків

Підготовка даних є першим і найдовшим етапом аналітичного прогнозування. Отримані дані можуть містити пропуски, помилки, неспівпадіння типів, форматів, зібраних даних може виявитися недостатньо або навіть забагато для аналізу, у випадку, якщо виявиться, що частина даних не вплине суттєво на результат прогнозування. Усі ці недоліки варто усунути перед етапом моделювання, щоб вони не сказалися на роботі моделі. Крім того, підготовка даних також може включати в себе нормалізацію даних і розбиття на тестову і робочу вибірки в залежності від обраних моделей.

Будемо проводити підготовку даних за допомогою SQL (Structured Query Language). Оскільки оригінальний файл даних не є таблицею бази даних SQL, а має розширення .csv, перетворимо його за допомогою засобів Microsoft Visual Studio:

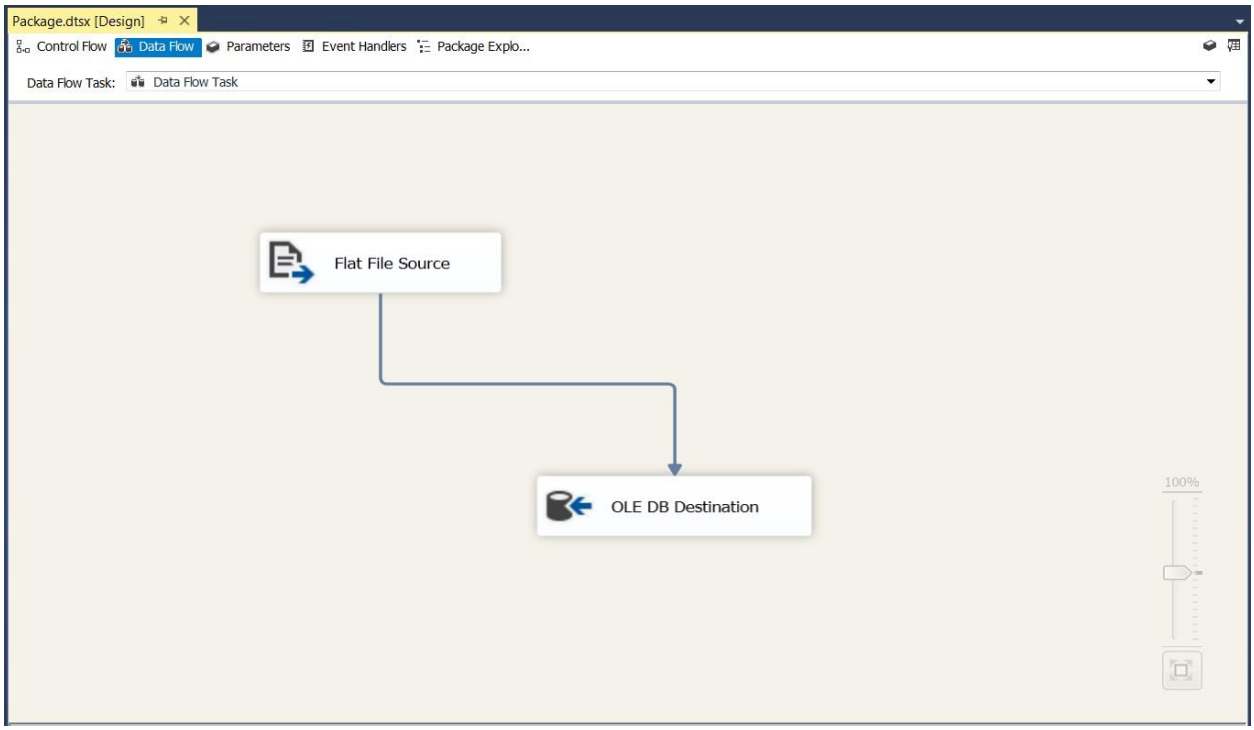
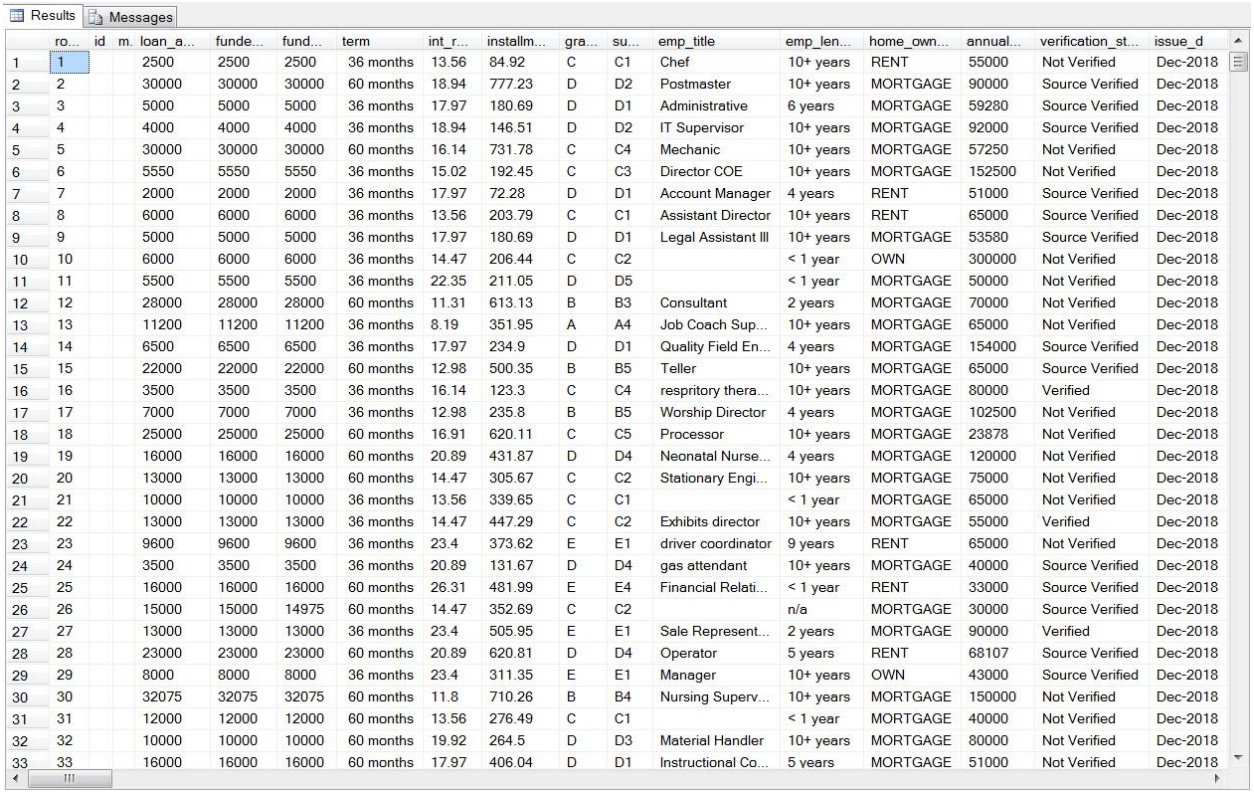



Рисунок 1 – Вигляд вікна Visual Studio

Отримуємо таблицю бази даних SQL. Проведемо огляд таблиці:



Results | Messages

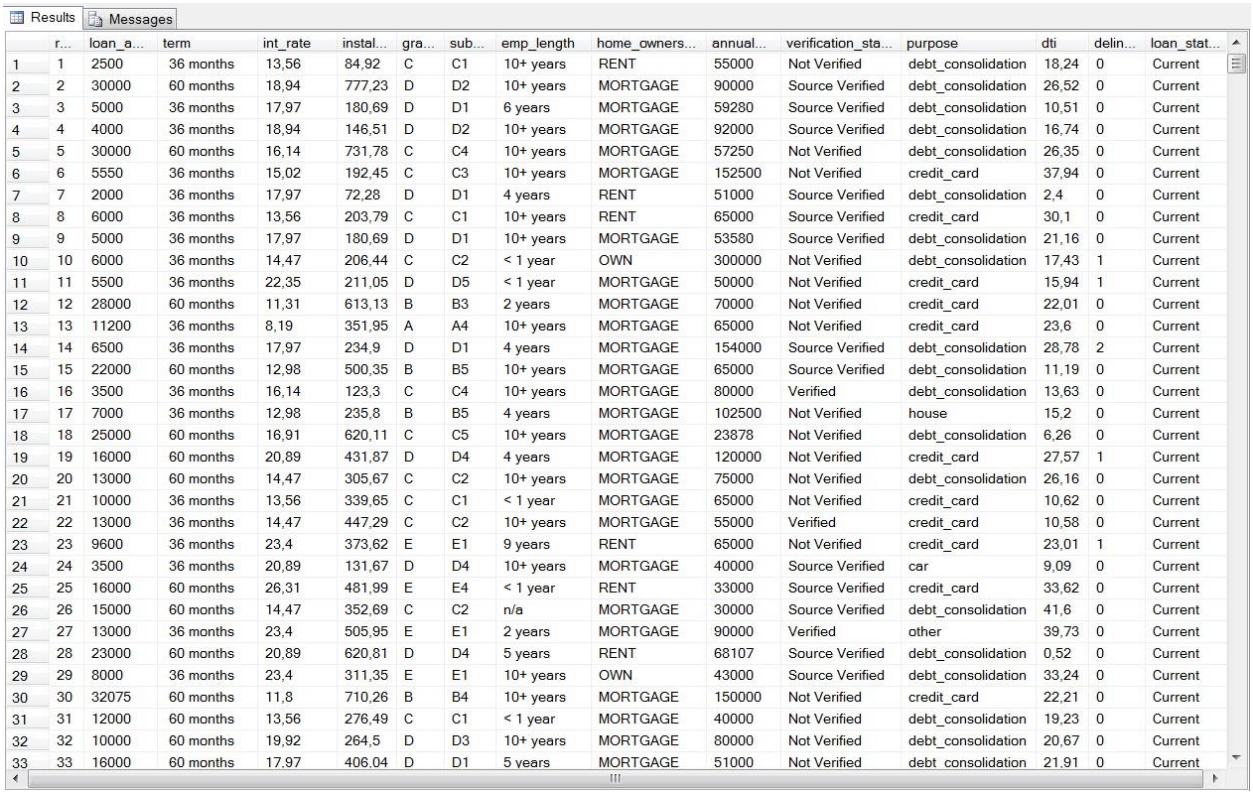
| | ro... | id | m. | loan_a... | funde... | fund... | term | int_r... | installm... | gra... | su... | emp_title | emp_len... | home_own... | annual... | verification_st... | issue_d |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | | | 2500 | 2500 | 2500 | 36 months | 13.56 | 84.92 | C | C1 | Chef | 10+ years | RENT | 55000 | Not Verified | Dec-2018 |
| 2 | 2 | | | 30000 | 30000 | 30000 | 60 months | 18.94 | 777.23 | D | D2 | Postmaster | 10+ years | MORTGAGE | 90000 | Source Verified | Dec-2018 |
| 3 | 3 | | | 5000 | 5000 | 5000 | 36 months | 17.97 | 180.69 | D | D1 | Administrative | 6 years | MORTGAGE | 59280 | Source Verified | Dec-2018 |
| 4 | 4 | | | 4000 | 4000 | 4000 | 36 months | 18.94 | 146.51 | D | D2 | IT Supervisor | 10+ years | MORTGAGE | 92000 | Source Verified | Dec-2018 |
| 5 | 5 | | | 30000 | 30000 | 30000 | 60 months | 16.14 | 731.78 | C | C4 | Mechanic | 10+ years | MORTGAGE | 57250 | Not Verified | Dec-2018 |
| 6 | 6 | | | 5550 | 5550 | 5550 | 36 months | 15.02 | 192.45 | C | C3 | Director COE | 10+ years | MORTGAGE | 152500 | Not Verified | Dec-2018 |
| 7 | 7 | | | 2000 | 2000 | 2000 | 36 months | 17.97 | 72.28 | D | D1 | Account Manager | 4 years | RENT | 51000 | Source Verified | Dec-2018 |
| 8 | 8 | | | 6000 | 6000 | 6000 | 36 months | 13.56 | 203.79 | C | C1 | Assistant Director | 10+ years | RENT | 65000 | Source Verified | Dec-2018 |
| 9 | 9 | | | 5000 | 5000 | 5000 | 36 months | 17.97 | 180.69 | D | D1 | Legal Assistant III | 10+ years | MORTGAGE | 53580 | Source Verified | Dec-2018 |
| 10 | 10 | | | 6000 | 6000 | 6000 | 36 months | 14.47 | 206.44 | C | C2 | | < 1 year | OWN | 300000 | Not Verified | Dec-2018 |
| 11 | 11 | | | 5500 | 5500 | 5500 | 36 months | 22.35 | 211.05 | D | D5 | | < 1 year | MORTGAGE | 50000 | Not Verified | Dec-2018 |
| 12 | 12 | | | 28000 | 28000 | 28000 | 60 months | 11.31 | 613.13 | B | B3 | Consultant | 2 years | MORTGAGE | 70000 | Not Verified | Dec-2018 |
| 13 | 13 | | | 11200 | 11200 | 11200 | 36 months | 8.19 | 351.95 | A | A4 | Job Coach Sup... | 10+ years | MORTGAGE | 65000 | Not Verified | Dec-2018 |
| 14 | 14 | | | 6500 | 6500 | 6500 | 36 months | 17.97 | 234.9 | D | D1 | Quality Field En... | 4 years | MORTGAGE | 154000 | Source Verified | Dec-2018 |
| 15 | 15 | | | 22000 | 22000 | 22000 | 60 months | 12.98 | 500.35 | B | B5 | Teller | 10+ years | MORTGAGE | 65000 | Source Verified | Dec-2018 |
| 16 | 16 | | | 3500 | 3500 | 3500 | 36 months | 16.14 | 123.3 | C | C4 | respritory thera... | 10+ years | MORTGAGE | 80000 | Verified | Dec-2018 |
| 17 | 17 | | | 7000 | 7000 | 7000 | 36 months | 12.98 | 235.8 | B | B5 | Worship Director | 4 years | MORTGAGE | 102500 | Not Verified | Dec-2018 |
| 18 | 18 | | | 25000 | 25000 | 25000 | 60 months | 16.91 | 620.11 | C | C5 | Processor | 10+ years | MORTGAGE | 23878 | Not Verified | Dec-2018 |
| 19 | 19 | | | 16000 | 16000 | 16000 | 60 months | 20.89 | 431.87 | D | D4 | Neonatal Nurse... | 4 years | MORTGAGE | 120000 | Not Verified | Dec-2018 |
| 20 | 20 | | | 13000 | 13000 | 13000 | 60 months | 14.47 | 305.67 | C | C2 | Stationary Engi... | 10+ years | MORTGAGE | 75000 | Not Verified | Dec-2018 |
| 21 | 21 | | | 10000 | 10000 | 10000 | 36 months | 13.56 | 339.65 | C | C1 | | < 1 year | MORTGAGE | 65000 | Not Verified | Dec-2018 |
| 22 | 22 | | | 13000 | 13000 | 13000 | 36 months | 14.47 | 447.29 | C | C2 | Exhibits director | 10+ years | MORTGAGE | 55000 | Verified | Dec-2018 |
| 23 | 23 | | | 9600 | 9600 | 9600 | 36 months | 23.4 | 373.62 | E | E1 | driver coordinator | 9 years | RENT | 65000 | Not Verified | Dec-2018 |
| 24 | 24 | | | 3500 | 3500 | 3500 | 36 months | 20.89 | 131.67 | D | D4 | gas attendant | 10+ years | MORTGAGE | 40000 | Source Verified | Dec-2018 |
| 25 | 25 | | | 16000 | 16000 | 16000 | 60 months | 26.31 | 481.99 | E | E4 | Financial Relati... | < 1 year | RENT | 33000 | Source Verified | Dec-2018 |
| 26 | 26 | | | 15000 | 15000 | 14975 | 60 months | 14.47 | 352.69 | C | C2 | | n/a | MORTGAGE | 30000 | Source Verified | Dec-2018 |
| 27 | 27 | | | 13000 | 13000 | 13000 | 36 months | 23.4 | 505.95 | E | E1 | Sale Represent... | 2 years | MORTGAGE | 90000 | Verified | Dec-2018 |
| 28 | 28 | | | 23000 | 23000 | 23000 | 60 months | 20.89 | 620.81 | D | D4 | Operator | 5 years | RENT | 68107 | Source Verified | Dec-2018 |
| 29 | 29 | | | 8000 | 8000 | 8000 | 36 months | 23.4 | 311.35 | E | E1 | Manager | 10+ years | OWN | 43000 | Source Verified | Dec-2018 |
| 30 | 30 | | | 32075 | 32075 | 32075 | 60 months | 11.8 | 710.26 | B | B4 | Nursing Superv... | 10+ years | MORTGAGE | 150000 | Not Verified | Dec-2018 |
| 31 | 31 | | | 12000 | 12000 | 12000 | 60 months | 13.56 | 276.49 | C | C1 | | < 1 year | MORTGAGE | 40000 | Not Verified | Dec-2018 |
| 32 | 32 | | | 10000 | 10000 | 10000 | 60 months | 19.92 | 264.5 | D | D3 | Material Handler | 10+ years | MORTGAGE | 80000 | Not Verified | Dec-2018 |
| 33 | 33 | | | 16000 | 16000 | 16000 | 60 months | 17.97 | 406.04 | D | D1 | Instructional Co... | 5 years | MORTGAGE | 51000 | Not Verified | Dec-2018 |

Рисунок 2 – Вигляд отриманої таблиці бази даних, аркуш 11

Таблиця має 153 поля (стовпців) і більше двох мільйонів записів. Значна кількість полів не суттєво впливатиме на результат передбачення. Виберемо необхідні поля і створимо с ними нову таблицю, з якою будемо надалі працювати.

Текст коду SQL із створенням та заповненням нової таблиці наведений у додатку А.

Маємо таблицю:



Results | Messages

| | r... | loan_a... | term | int_rate | instal... | gra... | sub... | emp_length | home_owners... | annual... | verification_sta... | purpose | dti | delin... | loan_stat... |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 2500 | 36 months | 13,56 | 84,92 | C | C1 | 10+ years | RENT | 55000 | Not Verified | debt_consolidation | 18,24 | 0 | Current |
| 2 | 2 | 30000 | 60 months | 18,94 | 777,23 | D | D2 | 10+ years | MORTGAGE | 90000 | Source Verified | debt_consolidation | 26,52 | 0 | Current |
| 3 | 3 | 5000 | 36 months | 17,97 | 180,69 | D | D1 | 6 years | MORTGAGE | 59280 | Source Verified | debt_consolidation | 10,51 | 0 | Current |
| 4 | 4 | 4000 | 36 months | 18,94 | 146,51 | D | D2 | 10+ years | MORTGAGE | 92000 | Source Verified | debt_consolidation | 16,74 | 0 | Current |
| 5 | 5 | 30000 | 60 months | 16,14 | 731,78 | C | C4 | 10+ years | MORTGAGE | 57250 | Not Verified | debt_consolidation | 26,35 | 0 | Current |
| 6 | 6 | 5550 | 36 months | 15,02 | 192,45 | C | C3 | 10+ years | MORTGAGE | 152500 | Not Verified | credit_card | 37,94 | 0 | Current |
| 7 | 7 | 2000 | 36 months | 17,97 | 72,28 | D | D1 | 4 years | RENT | 51000 | Source Verified | debt_consolidation | 2,4 | 0 | Current |
| 8 | 8 | 6000 | 36 months | 13,56 | 203,79 | C | C1 | 10+ years | RENT | 65000 | Source Verified | credit_card | 30,1 | 0 | Current |
| 9 | 9 | 5000 | 36 months | 17,97 | 180,69 | D | D1 | 10+ years | MORTGAGE | 53580 | Source Verified | debt_consolidation | 21,16 | 0 | Current |
| 10 | 10 | 6000 | 36 months | 14,47 | 206,44 | C | C2 | < 1 year | OWN | 300000 | Not Verified | debt_consolidation | 17,43 | 1 | Current |
| 11 | 11 | 5500 | 36 months | 22,35 | 211,05 | D | D5 | < 1 year | MORTGAGE | 50000 | Not Verified | credit_card | 15,94 | 1 | Current |
| 12 | 12 | 28000 | 60 months | 11,31 | 613,13 | B | B3 | 2 years | MORTGAGE | 70000 | Not Verified | credit_card | 22,01 | 0 | Current |
| 13 | 13 | 11200 | 36 months | 8,19 | 351,95 | A | A4 | 10+ years | MORTGAGE | 65000 | Not Verified | credit_card | 23,6 | 0 | Current |
| 14 | 14 | 6500 | 36 months | 17,97 | 234,9 | D | D1 | 4 years | MORTGAGE | 154000 | Source Verified | debt_consolidation | 28,78 | 2 | Current |
| 15 | 15 | 22000 | 60 months | 12,98 | 500,35 | B | B5 | 10+ years | MORTGAGE | 65000 | Source Verified | debt_consolidation | 11,19 | 0 | Current |
| 16 | 16 | 3500 | 36 months | 16,14 | 123,3 | C | C4 | 10+ years | MORTGAGE | 80000 | Verified | debt_consolidation | 13,63 | 0 | Current |
| 17 | 17 | 7000 | 36 months | 12,98 | 235,8 | B | B5 | 4 years | MORTGAGE | 102500 | Not Verified | house | 15,2 | 0 | Current |
| 18 | 18 | 25000 | 60 months | 16,91 | 620,11 | C | C5 | 10+ years | MORTGAGE | 23878 | Not Verified | debt_consolidation | 6,26 | 0 | Current |
| 19 | 19 | 16000 | 60 months | 20,89 | 431,87 | D | D4 | 4 years | MORTGAGE | 120000 | Not Verified | credit_card | 27,57 | 1 | Current |
| 20 | 20 | 13000 | 60 months | 14,47 | 305,67 | C | C2 | 10+ years | MORTGAGE | 75000 | Not Verified | debt_consolidation | 26,16 | 0 | Current |
| 21 | 21 | 10000 | 36 months | 13,56 | 339,65 | C | C1 | < 1 year | MORTGAGE | 65000 | Not Verified | credit_card | 10,62 | 0 | Current |
| 22 | 22 | 13000 | 36 months | 14,47 | 447,29 | C | C2 | 10+ years | MORTGAGE | 55000 | Verified | credit_card | 10,58 | 0 | Current |
| 23 | 23 | 9600 | 36 months | 23,4 | 373,62 | E | E1 | 9 years | RENT | 65000 | Not Verified | credit_card | 23,01 | 1 | Current |
| 24 | 24 | 3500 | 36 months | 20,89 | 131,67 | D | D4 | 10+ years | MORTGAGE | 40000 | Source Verified | car | 9,09 | 0 | Current |
| 25 | 25 | 16000 | 60 months | 26,31 | 481,99 | E | E4 | < 1 year | RENT | 33000 | Source Verified | credit_card | 33,62 | 0 | Current |
| 26 | 26 | 15000 | 60 months | 14,47 | 352,69 | C | C2 | n/a | MORTGAGE | 30000 | Source Verified | debt_consolidation | 41,6 | 0 | Current |
| 27 | 27 | 13000 | 36 months | 23,4 | 505,95 | E | E1 | 2 years | MORTGAGE | 90000 | Verified | other | 39,73 | 0 | Current |
| 28 | 28 | 23000 | 60 months | 20,89 | 620,81 | D | D4 | 5 years | RENT | 68107 | Source Verified | debt_consolidation | 0,52 | 0 | Current |
| 29 | 29 | 8000 | 36 months | 23,4 | 311,35 | E | E1 | 10+ years | OWN | 43000 | Source Verified | debt_consolidation | 33,24 | 0 | Current |
| 30 | 30 | 32075 | 60 months | 11,8 | 710,26 | B | B4 | 10+ years | MORTGAGE | 150000 | Not Verified | credit_card | 22,21 | 0 | Current |
| 31 | 31 | 12000 | 60 months | 13,56 | 276,49 | C | C1 | < 1 year | MORTGAGE | 40000 | Not Verified | debt_consolidation | 19,23 | 0 | Current |
| 32 | 32 | 10000 | 60 months | 19,92 | 264,5 | D | D3 | 10+ years | MORTGAGE | 80000 | Not Verified | debt_consolidation | 20,67 | 0 | Current |
| 33 | 33 | 16000 | 60 months | 17,97 | 406,04 | D | D1 | 5 years | MORTGAGE | 51000 | Not Verified | debt_consolidation | 21,91 | 0 | Current |

Рисунок 3, аркуш 12

Список полів таблиці:

1. row_number – номер рядка
2. loan_amnt – loan amount, сума кредиту;
3. term – термін кредиту;
4. int_rate – interest rate, відсоткова ставка;
5. installment – щомісячний платіж;
6. grade – клас кредиту (A-G);
7. sub_grade – субклас кредиту (A1-G5);
8. emp_length – employment length, термін працевлаштування клієнта;
9. home_ownership – наявність приватного житла клієнтом;
10. annual_inc – annual income, річний дохід клієнта;
11. verification_status – наявність верифікації рівня доходу клієнта компанією;
12. purpose – ціль взяття кредиту;

13. dti – debt-to-income, числова величина, що дорівнює відношенню щомісячних виплат клієта по усім його кредитам до його щомісячного доходу;
14. delinq_2yrs – кількість випадків правопорушення у кредитній справі клієнта за останні 2 роки;
15. loan_status – поточний статус кредиту.

Отримана таблиця має 2 265 101 записів.

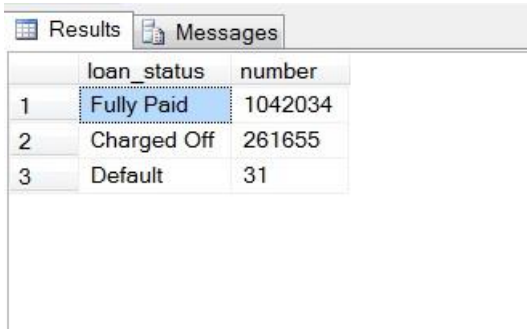
Отримаємо картину розподілу статусу кредів:

Results | Messages

| | loan_status | num |
|---|---|---|
| 1 | Fully Paid | 1042034 |
| 2 | Current | 924042 |
| 3 | Charged Off | 261655 |
| 4 | Late (31-120 days) | 21901 |
| 5 | In Grace Period | 8953 |
| 6 | Late (16-30 days) | 3737 |
| 7 | Does not meet the credit policy. Status:Fully Paid | 1988 |
| 8 | Does not meet the credit policy. Status:Charged ... | 760 |
| 9 | Default | 31 |

Рисунок 4 – Розподіл статусів кредитів

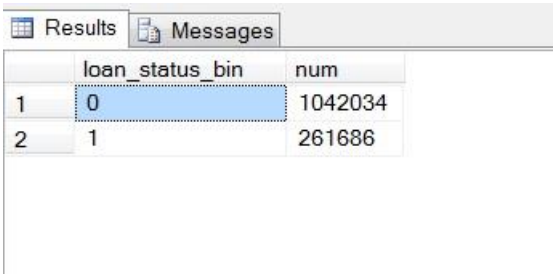
Як бачимо, трохи менше половини кредитів мають статус «Current» (поточний), результат цих кредитів ще невідомий. Також є кредити зі статусом запізнення на різний термін та інші. Ці записи необхідно видалити з таблиці:




| | loan_status | number |
|---|---|---|
| 1 | Fully Paid | 1042034 |
| 2 | Charged Off | 261655 |
| 3 | Default | 31 |

Рисунок 5, аркуш 14 – Розподіл статусів після видалення непотрібних

Будемо класифікувати погашені та непогашені (дефолтні) кредити. Для цього позначимо записи зі статусом кредиту «Fully Paid» як 0, зі статусами «Charged Off» та «Default» як 1:




| | loan_status_bin | num |
|---|---|---|
| 1 | 0 | 1042034 |
| 2 | 1 | 261686 |

Рисунок 6

Текст коду SQL із цим перетворенням таблиці наведений у додатку А.

## 1.2 Основні характеристики даних щодо кредитних випадків

Продублюємо рис. 16 у вигляді гістограми:

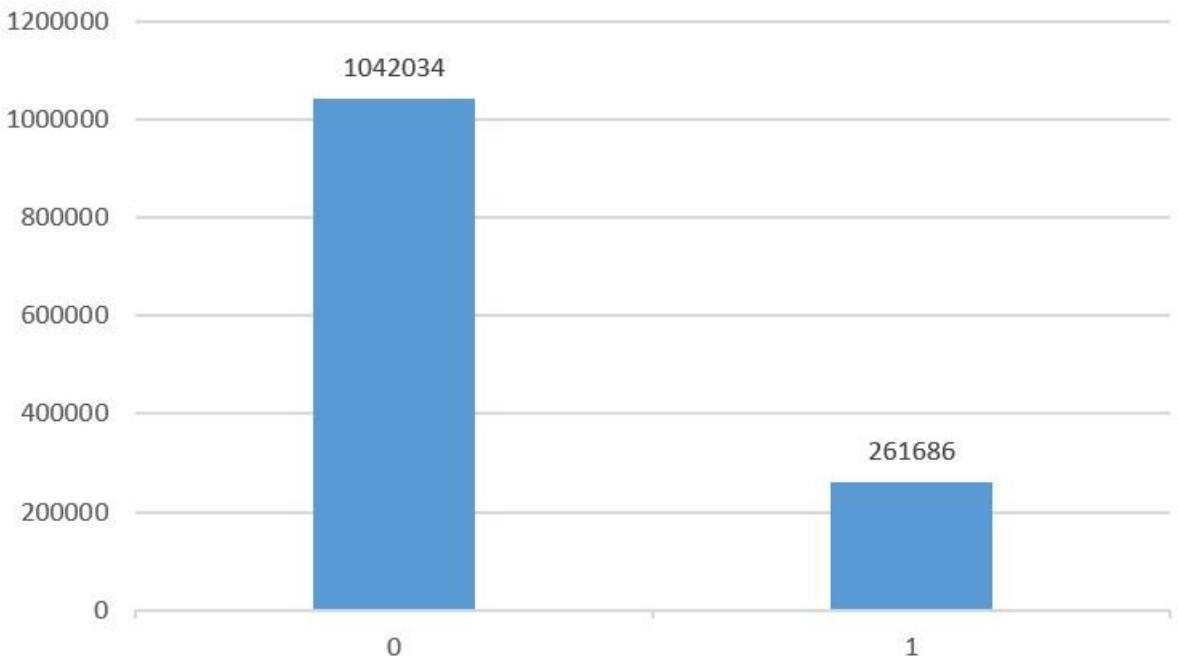



Рисунок 7 – Гістограма статусу кредитів (loan_status)

Як бачимо на рис.17, кількість погашених кредитів (0) суттєво переважує кількість дефолтних (1) – дані незбалансовані.

Також приведемо гістограми і по іншим полям даних:

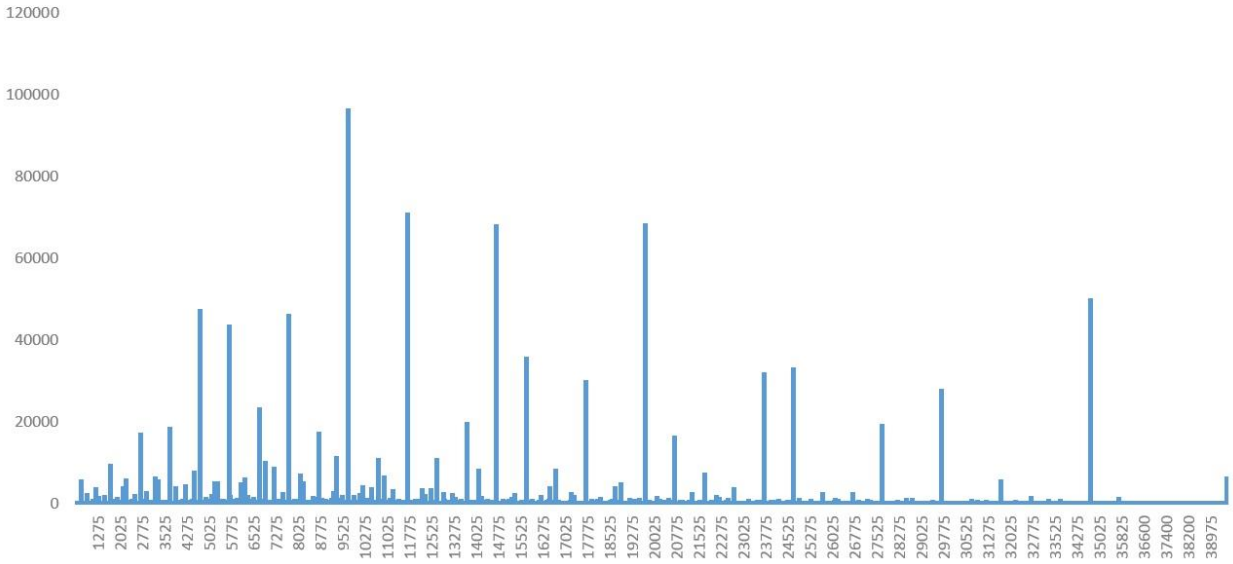



Рисунок 8 – Розмір кредитів (loan_amnt)

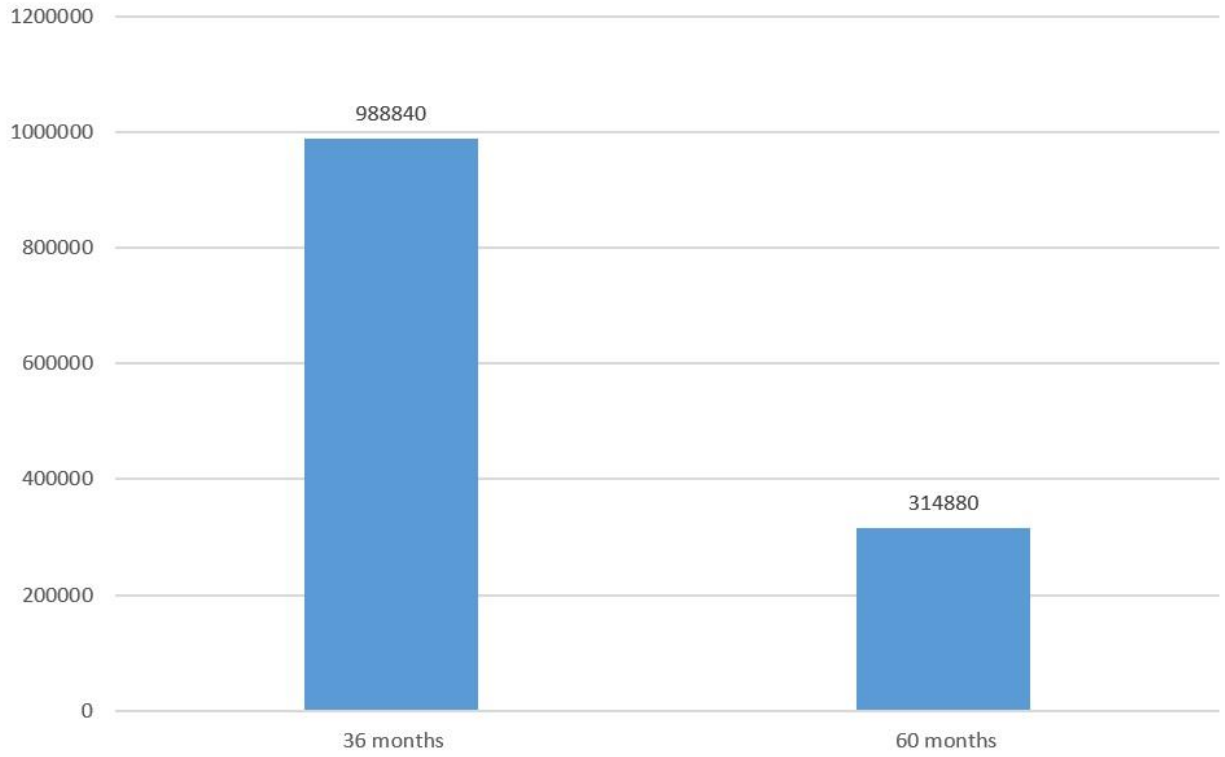



Рисунок 9 – Терміни кредитів (term)

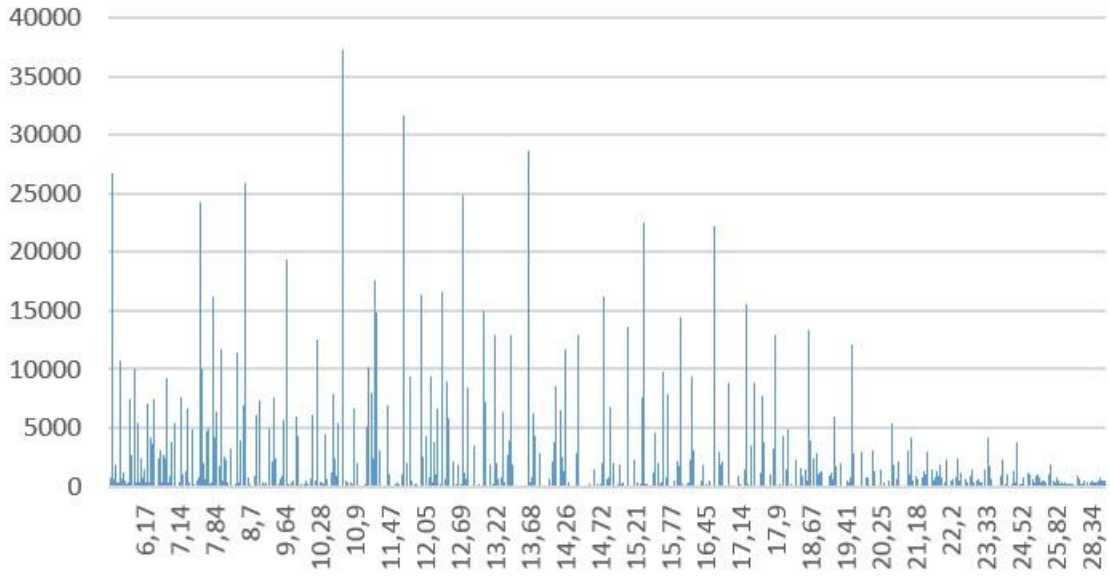



Рисунок 10 – Відсоткова ставка кредитів (int_rate)

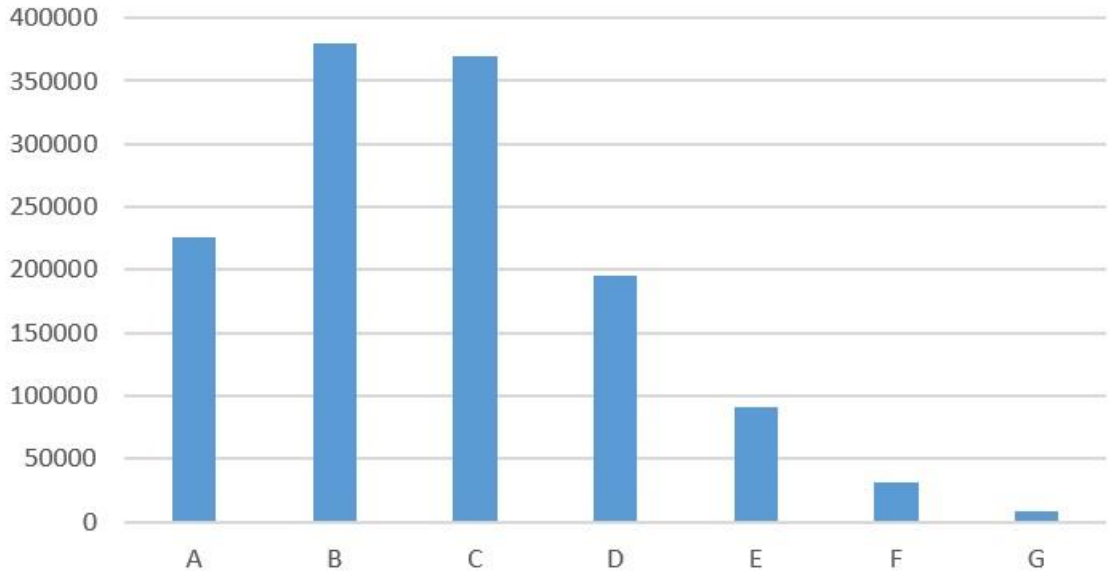



Рисунок 11 – Класи кредитів (grade)

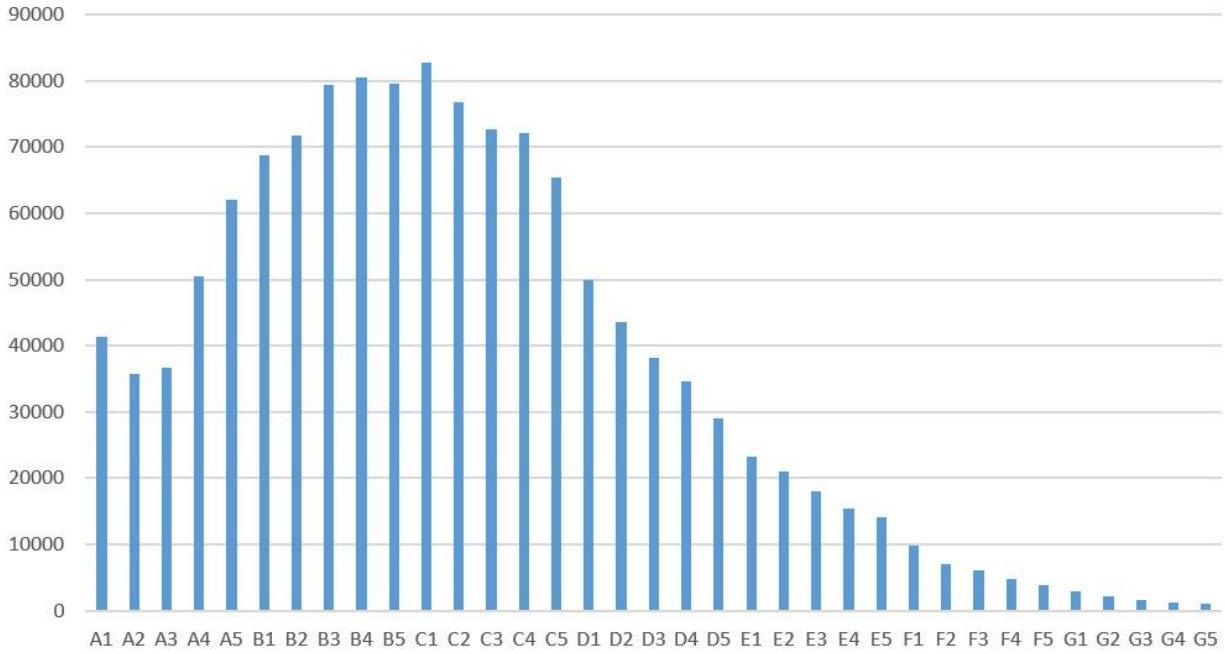



Рисунок 12 – Субкласи кредитів (sub_grade)

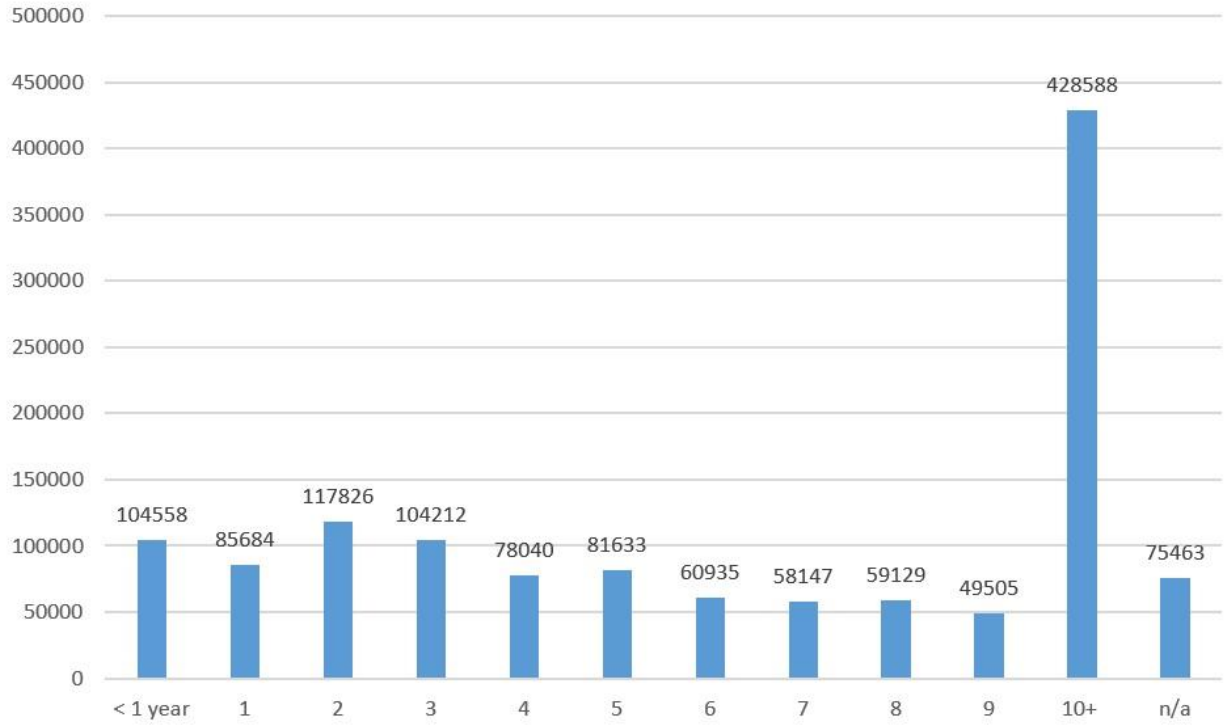



Рисунок 13 – Термін працевлаштування клієнта у роках (emp_length)


Рисунок 14 – Тип житла клієнта (home_ownership), (MORTGAGE – іпотека, OWN – власність, RENT – оренда, ANY – будь який, NONE – без житла, OTHER – інші)


Рисунок 15 - Наявність верифікації рівня доходу клієнта компанією (verification_status), (Not Verified – не верифіковано, Verified – верифіковано, Source Verified – верифіковано джерелом)

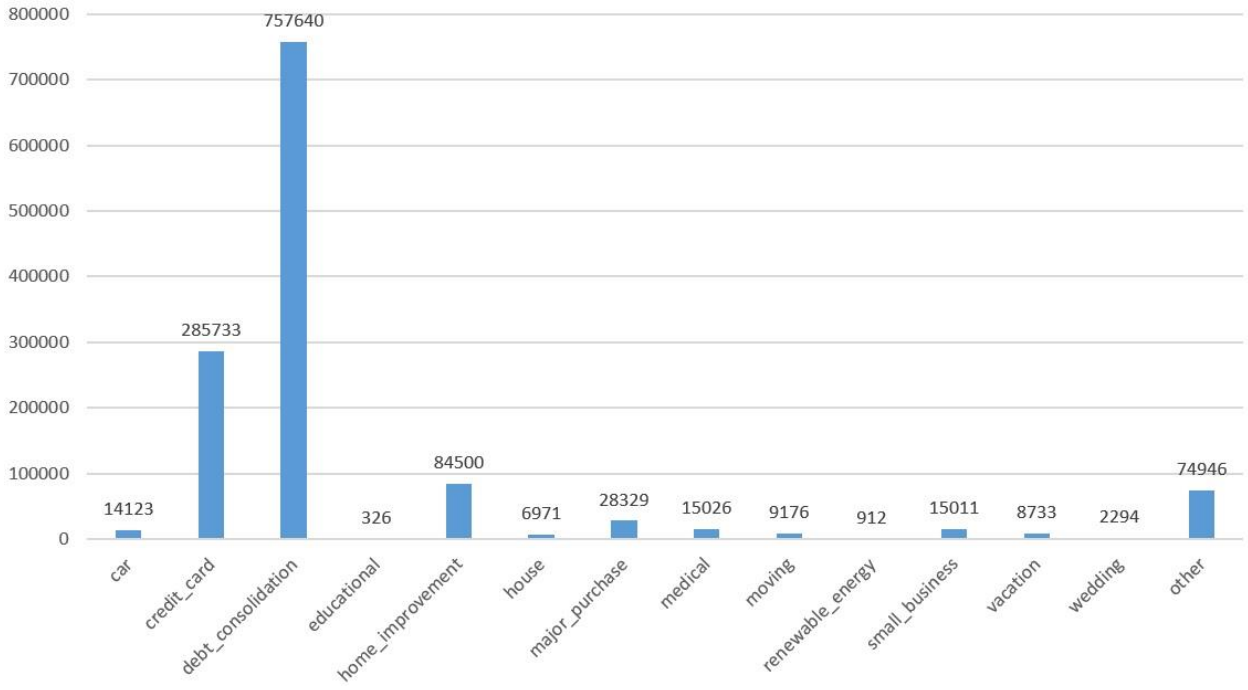



Рисунок 16 – ціль кредиту (purpose), (car – авто, credit_card – кредитний рахунок, debt_consolidation – рефінансування заборгованості, educational – навчання, home_improvement – покращення домівки, house – будинок, major_purchase – велика покупка, medical – лікування, moving – переїзд, renewable_energy - відновлювальна енергія, small_business – малий бізнес, vacation – відпочинок, wedding – весілля, other – інші)

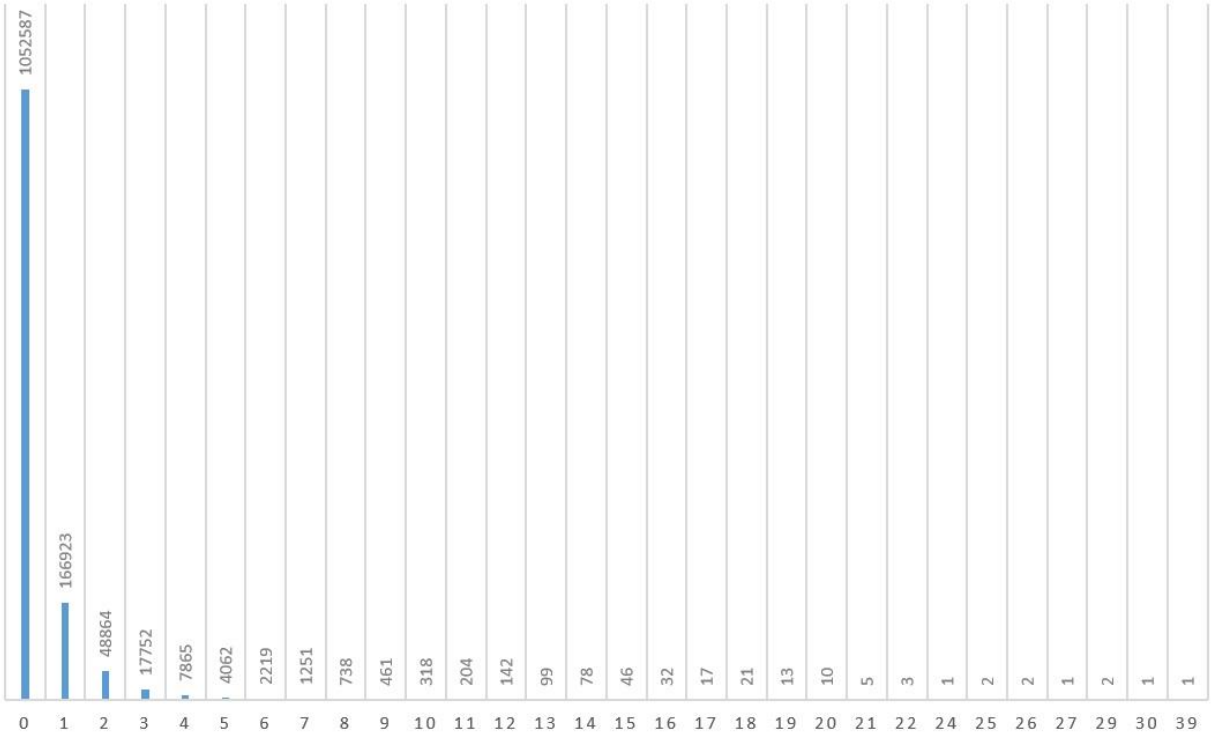



Рисунок 17 - Кількість випадків правопорушення у кредитній справі клієнта за останні 2 роки

Побудуємо кореляційну матрицю полів:

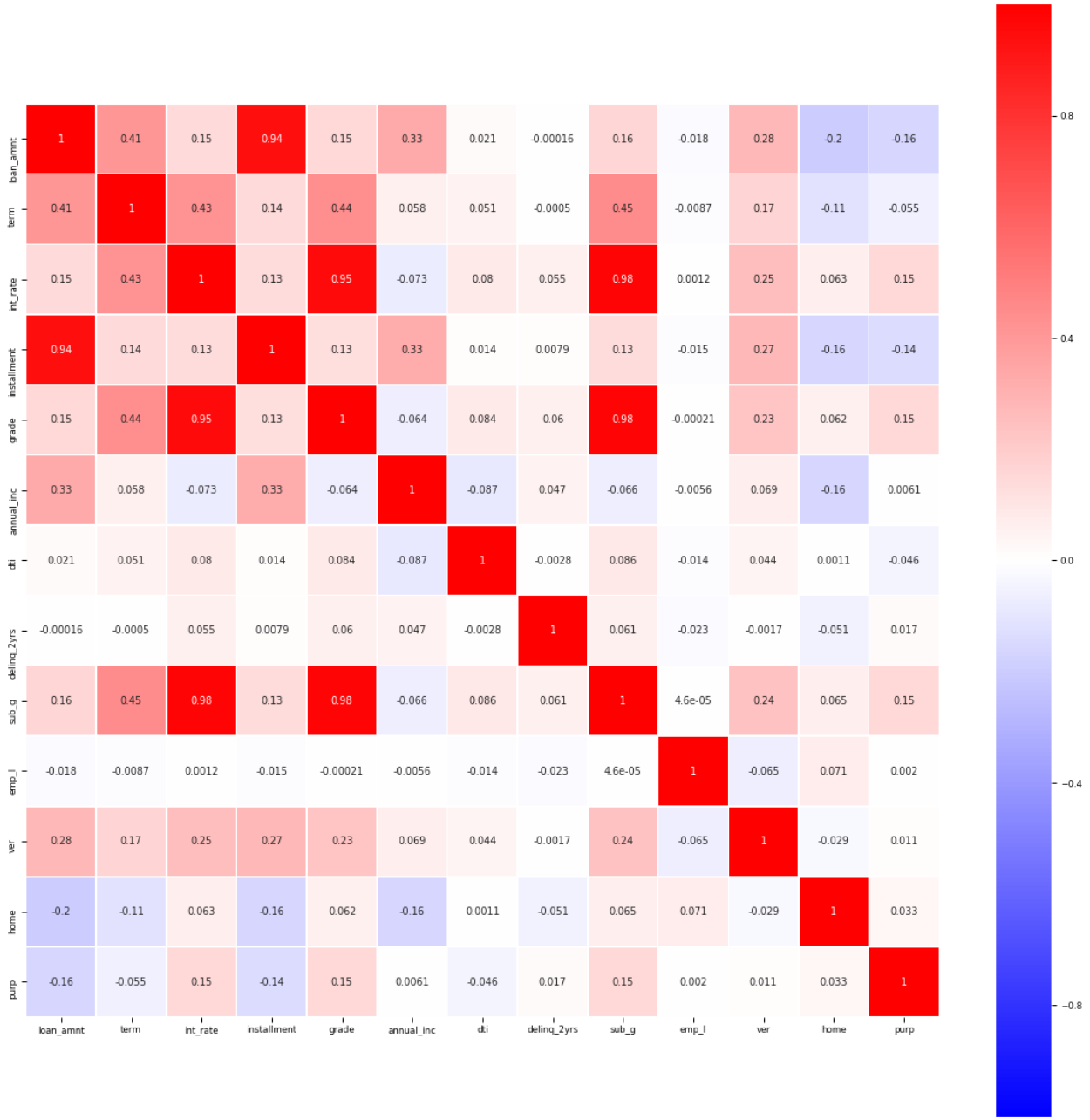



Рисунок 18 – Кореляційна матриця полів

Як бачимо, щомісячний платіж (installment) і розмір кредиту (loan_amnt) мають високу кореляцію (0.94), що є логічним. Клас кредиту (grade) і відсоткова ставка (int_rate) мають кореляцію 0.95. Субклас кредиту (sub_grade) і відсоткова ставка мають кореляцію 0.98. Річ у тім, що відсоткова ставка обирається напряму в залежності від класу кредиту. Видалимо поля

installment, grade і sub_grade на користь loan_amnt і int_rate щоб запобігти мультиколінеарності.

Крім того, гістограма рис. 21 показує, що є чимала кількість кредитів, у яких є невідомим термін працевлаштування клієнта (n/a). Видалимо записи з такою характеристикою.

Текст коду SQL із цим перетворенням таблиці наведений у додатку А.

Отримана таблиця має вигляд:

Results | Messages

| | row_number | loan_amnt | term | int_rate | emp_length | home_ownership | annual_inc | verification_status | purpose | dti | delinq_2yrs | loan_status |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 12000 | 60 months | 15,99 | 10+ years | RENT | 86400 | Source Verified | car | 8,94 | 0 | 1 |
| 2 | 2 | 20000 | 36 months | 7,49 | 7 years | OWN | 55000 | Verified | other | 10,89 | 0 | 0 |
| 3 | 3 | 9000 | 36 months | 18,39 | 4 years | RENT | 42000 | Source Verified | debt_consolidation | 18,23 | 1 | 1 |
| 4 | 4 | 13250 | 60 months | 16,49 | 1 year | RENT | 38400 | Not Verified | debt_consolidation | 23,66 | 1 | 0 |
| 5 | 5 | 10000 | 36 months | 13,99 | < 1 year | MORTGAGE | 285000 | Verified | major_purchase | 14,27 | 0 | 0 |
| 6 | 6 | 10000 | 36 months | 5,99 | < 1 year | RENT | 45500 | Source Verified | debt_consolidation | 22,15 | 0 | 0 |
| 7 | 7 | 2000 | 60 months | 20,25 | 2 years | RENT | 39000 | Source Verified | other | 8,89 | 0 | 0 |
| 8 | 8 | 12000 | 60 months | 20,25 | 2 years | RENT | 76000 | Verified | small_business | 1,66 | 0 | 0 |
| 9 | 9 | 4000 | 36 months | 10,99 | 2 years | MORTGAGE | 98004 | Not Verified | other | 9,48 | 0 | 0 |
| 10 | 10 | 35000 | 60 months | 19,69 | 10+ years | RENT | 130000 | Source Verified | debt_consolidation | 4,53 | 0 | 0 |
| 11 | 11 | 34800 | 60 months | 20,11 | < 1 year | RENT | 120000 | Source Verified | debt_consolidation | 8,38 | 0 | 0 |
| 12 | 12 | 3200 | 36 months | 10,99 | 3 years | MORTGAGE | 40000 | Not Verified | debt_consolidation | 13,17 | 0 | 0 |
| 13 | 13 | 14000 | 36 months | 10,99 | 4 years | OWN | 71000 | Verified | debt_consolidation | 23,35 | 0 | 0 |
| 14 | 14 | 18000 | 36 months | 8,49 | 10+ years | RENT | 73500 | Not Verified | credit_card | 14,68 | 0 | 0 |
| 15 | 15 | 4000 | 36 months | 9,99 | 2 years | RENT | 26280 | Source Verified | debt_consolidation | 23,29 | 0 | 0 |
| 16 | 16 | 7350 | 36 months | 6,92 | 1 year | RENT | 12480 | Not Verified | small_business | 1,54 | 0 | 0 |
| 17 | 17 | 17000 | 60 months | 11,11 | 10+ years | MORTGAGE | 59000 | Not Verified | other | 3,42 | 0 | 0 |
| 18 | 18 | 10000 | 60 months | 10,37 | 10+ years | OWN | 60000 | Not Verified | home_improvem... | 24,08 | 0 | 1 |
| 19 | 19 | 8000 | 36 months | 9,63 | 5 years | RENT | 68848 | Not Verified | debt_consolidation | 10,32 | 0 | 0 |
| 20 | 20 | 3500 | 36 months | 11,11 | 3 years | RENT | 74880 | Source Verified | car | 21,55 | 0 | 0 |
| 21 | 21 | 6000 | 60 months | 15,65 | 8 years | MORTGAGE | 50004 | Source Verified | debt_consolidation | 2,62 | 0 | 0 |
| 22 | 22 | 2500 | 36 months | 6,92 | 6 years | OWN | 31577 | Not Verified | car | 8,21 | 0 | 0 |
| 23 | 23 | 9600 | 60 months | 17,51 | 3 years | RENT | 65000 | Source Verified | small_business | 2,71 | 0 | 1 |
| 24 | 24 | 10000 | 36 months | 10 | 5 years | RENT | 29000 | Source Verified | debt_consolidation | 18,7 | 0 | 0 |
| 25 | 25 | 12000 | 60 months | 18,62 | < 1 year | RENT | 39900 | Verified | credit_card | 8,87 | 0 | 0 |
| 26 | 26 | 10000 | 60 months | 18,25 | 4 years | RENT | 36000 | Source Verified | small_business | 9 | 1 | 0 |
| 27 | 27 | 25000 | 36 months | 10,37 | 9 years | MORTGAGE | 150000 | Verified | debt_consolidation | 10,84 | 0 | 0 |
| 28 | 28 | 14000 | 36 months | 10,37 | 10+ years | RENT | 57600 | Source Verified | debt_consolidation | 20,02 | 0 | 1 |
| 29 | 29 | 10000 | 60 months | 19,36 | 7 years | MORTGAGE | 65000 | Not Verified | debt_consolidation | 6,65 | 1 | 1 |
| 30 | 30 | 18000 | 36 months | 11,11 | 3 years | MORTGAGE | 95000 | Source Verified | debt_consolidation | 16,02 | 0 | 0 |
| 31 | 31 | 4000 | 60 months | 13,43 | < 1 year | RENT | 18720 | Not Verified | moving | 7,69 | 0 | 0 |
| 32 | 32 | 8000 | 36 months | 9,63 | 6 years | RENT | 60000 | Verified | credit_card | 14,38 | 0 | 0 |
| 33 | 33 | 7000 | 36 months | 10,74 | 2 years | RENT | 75000 | Verified | other | 15,94 | 3 | 0 |
| 34 | 34 | 20000 | 36 months | 12,68 | 7 years | RENT | 62400 | Verified | debt_consolidation | 22 | 0 | 0 |

Рисунок 19

Ця таблиця має 1 228 257 записів.

Нова гістограма поля emp_length:

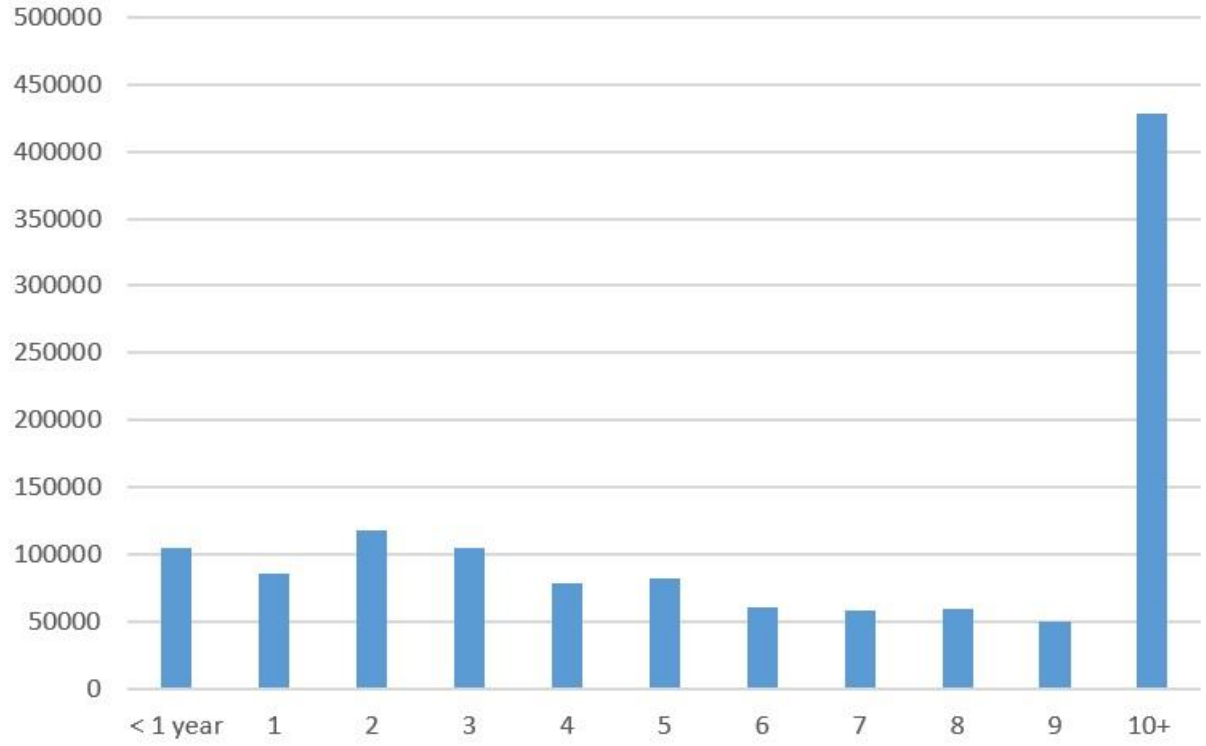



Рисунок 20, аркуш 24 – Термін працевлаштування клієнта у роках після видалення невідомих значень (emp_length)

Перетворимо таблицю назад у файл .csv для її використання у середовищі Python. Завершимо підготовку даних кодуванням категорійних ознак, розбиттям на тестову і робочу вибірки і нормалізацією.

Відкриємо файл і розіб’ємо його на описові ознаки (X) і цільову ознаку (y):

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 12000 | 1 | 15.99 | 1 | RENT | 86400.0 | Source Verified | car | 8.93999999999… | 0 |
| 1 | 20000 | 0 | 7.49 | 7 | OWN | 55000.0 | Verified | other | 10.89 | 0 |
| 2 | 9000 | 0 | 18.39 | 4 | RENT | 42000.0 | Source Verified | debt_consolid… | 18.23 | 1 |
| 3 | 13250 | 1 | 16.49 | 0 | RENT | 38400.0 | Not Verified | debt_consolid… | 23.66 | 1 |
| 4 | 10000 | 0 | 13.99 | 10 | MORTGAGE | 285000.0 | Verified | major_purchase | 14.27 | 0 |
| 5 | 10000 | 0 | 5.99 | 10 | RENT | 45500.0 | Source Verified | debt_consolid… | 22.15 | 0 |
| 6 | 2000 | 1 | 20.25 | 2 | RENT | 39000.0 | Source Verified | other | 8.89 | 0 |
| 7 | 12000 | 1 | 20.25 | 2 | RENT | 76000.0 | Verified | small_business | 1.66 | 0 |
| 8 | 4000 | 0 | 10.99 | 2 | MORTGAGE | 98004.0 | Not Verified | other | 9.48 | 0 |
| 9 | 35000 | 1 | 19.69 | 1 | RENT | 130000.0 | Source Verified | debt_consolid… | 4.53 | 0 |
| 10 | 34800 | 1 | 20.11 | 10 | RENT | 120000.0 | Source Verified | debt_consolid… | 8.38 | 0 |
| 11 | 3200 | 0 | 10.99 | 3 | MORTGAGE | 40000.0 | Not Verified | debt_consolid… | 13.17 | 0 |
| 12 | 14000 | 0 | 10.99 | 4 | OWN | 71000.0 | Verified | debt_consolid… | 23.35 | 0 |
| 13 | 18000 | 0 | 8.49 | 1 | RENT | 73500.0 | Not Verified | credit_card | 14.68 | 0 |
| 14 | 4000 | 0 | 9.99 | 2 | RENT | 26280.0 | Source Verified | debt_consolid… | 23.29 | 0 |
| 15 | 7350 | 0 | 6.92 | 0 | RENT | 12480.0 | Not Verified | small_business | 1.54 | 0 |
| 16 | 17000 | 1 | 11.11 | 1 | MORTGAGE | 59000.0 | Not Verified | other | 3.42 | 0 |
| 17 | 10000 | 1 | 10.37 | 1 | OWN | 60000.0 | Not Verified | home_improvem… | 24.08 | 0 |

Рисунок 21, аркуш 25

Ці ознаки містять порівнювальні категорії (термін кредиту 36 і 60 місяців, термін роботи від 1 року до 10 років і більше), то ж їх можна закодувати однією колонкою (колонки 1 і 3 на рис. 30)

З іншого боку, ознаки home_ownership, verification_status і purpose містять непорівнювальні категорії (наприклад, оренда житла чи власність, ціль кредиту автомобіль чи малий бізнес та ін.), тому ці ознаки варто перетворити на фіктивні змінні, як було розглянуто у розділі 1. Для кожної категорії створимо нову колонку, належність до якої позначимо нулем чи одиницею. Розіб’ємо дані на робочу (75%) і тестову (25%) частини. Після цього нормалізуємо дані (рис. 31).

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| 0 | 1.00592 | -0.00642282 | -0.0107781 | -0.336472 | -0.81799 | 1.50113 | -0.813833 |
| 1 | 1.00592 | -0.00642282 | -0.0107781 | -0.336472 | -0.81799 | -0.666166 | -0.813833 |
| 2 | 1.00592 | -0.00642282 | -0.0107781 | -0.336472 | -0.81799 | -0.666166 | 1.22875 |
| 3 | -0.994118 | -0.00642282 | -0.0107781 | -0.336472 | 1.22251 | -0.666166 | 1.22875 |
| 4 | 1.00592 | -0.00642282 | -0.0107781 | -0.336472 | -0.81799 | -0.666166 | -0.813833 |
| 5 | 1.00592 | -0.00642282 | -0.0107781 | -0.336472 | -0.81799 | -0.666166 | -0.813833 |
| 6 | -0.994118 | -0.00642282 | -0.0107781 | -0.336472 | 1.22251 | -0.666166 | 1.22875 |
| 7 | -0.994118 | -0.00642282 | -0.0107781 | -0.336472 | 1.22251 | -0.666166 | -0.813833 |
| 8 | -0.994118 | -0.00642282 | -0.0107781 | 2.97201 | -0.81799 | -0.666166 | 1.22875 |
| 9 | -0.994118 | -0.00642282 | -0.0107781 | -0.336472 | 1.22251 | -0.666166 | -0.813833 |
| 10 | -0.994118 | -0.00642282 | -0.0107781 | -0.336472 | 1.22251 | 1.50113 | -0.813833 |
| 11 | 1.00592 | -0.00642282 | -0.0107781 | -0.336472 | -0.81799 | -0.666166 | -0.813833 |
| 12 | 1.00592 | -0.00642282 | -0.0107781 | -0.336472 | -0.81799 | -0.666166 | 1.22875 |
| 13 | -0.994118 | -0.00642282 | -0.0107781 | -0.336472 | 1.22251 | -0.666166 | 1.22875 |

Рисунок 22, аркуш 26

Висновки до розділу 1 та постановка задачі дослідження

Процес підготовки даних є вкрай важливим етапом аналітичного прогнозування, адже від якості даних напряму залежить точність створеної моделі машинного навчання на основі цих даних. У процесі підготовки була виявлена особливість даних, яка полягає у тому, що дефолтні кредити становлять лише чверть усіх записів таблиці. Це необхідно брати до уваги в подальшому при виборі і створенні моделей машинного навчання.

Постановка задачі дослідження: провести аналіз даних організації-кредитора, вибрати моделі машинного навчання для класифікації кредитів. Виконати порівняльний аналіз обраних моделей. Спроектувати і реалізувати систему прийняття рішень для прогнозування результату кредитів.

# РОЗДІЛ 2. ОСНОВНІ АЛГОРИТМИ МАШИННОГО НАВЧАННЯ

Найбільш уживані моделі машинного навчання для прогнозування включають в себе:

- регресії (проста лінійна, лінійна, логістична);
- дерева рішень;
- наївна модель Байєса;
- алгоритм k-NN;

Розглянемо кожну з них.

## 2.1 Регресії

Проста лінійна регресія визначається функцією

$$y = b_0 + b_1 x_1 \quad (1)$$

де $y$ - залежна змінна,

$x_1$ - незалежна змінна,

$b_0$ і $b_1$ – коефіцієнти.

Наприклад, нам необхідно встановити залежність між віком працівника і його очікуваною заробітною платою. У ролі незалежної змінної буде виступати вік працівника, залежної змінної — заробітна плата. Графік простої лінійної регресії у цьому випадку може виглядати так:

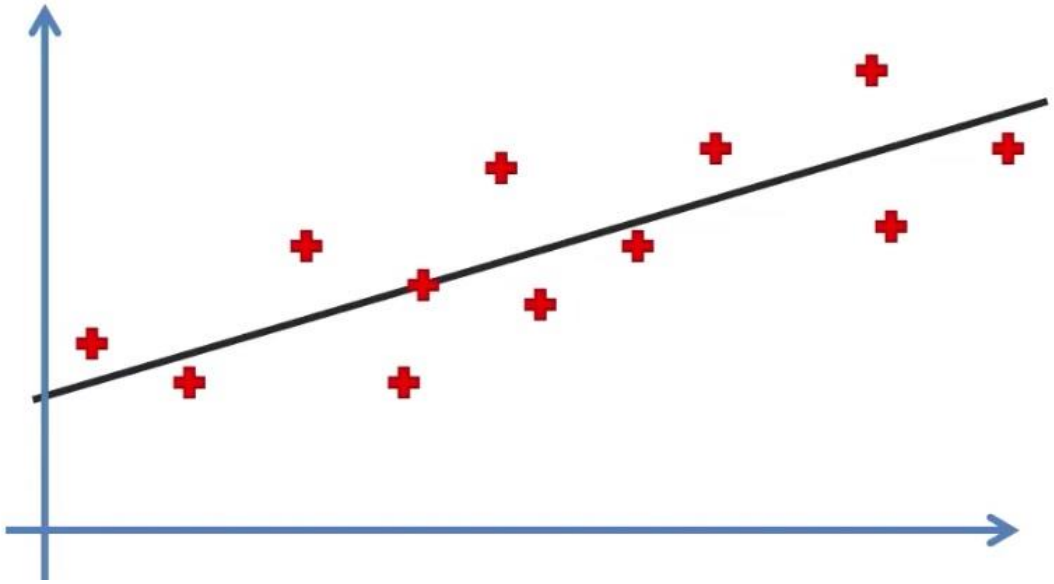


Рисунок 23, аркуш 28

На осі абсцис  позначено вік, а на осі ординат позначена заробітна плата. Коефіцієнти $b_0$ і $b_1$ підбираються таким чином, щоб сума квадратів похибок була мінімальною:

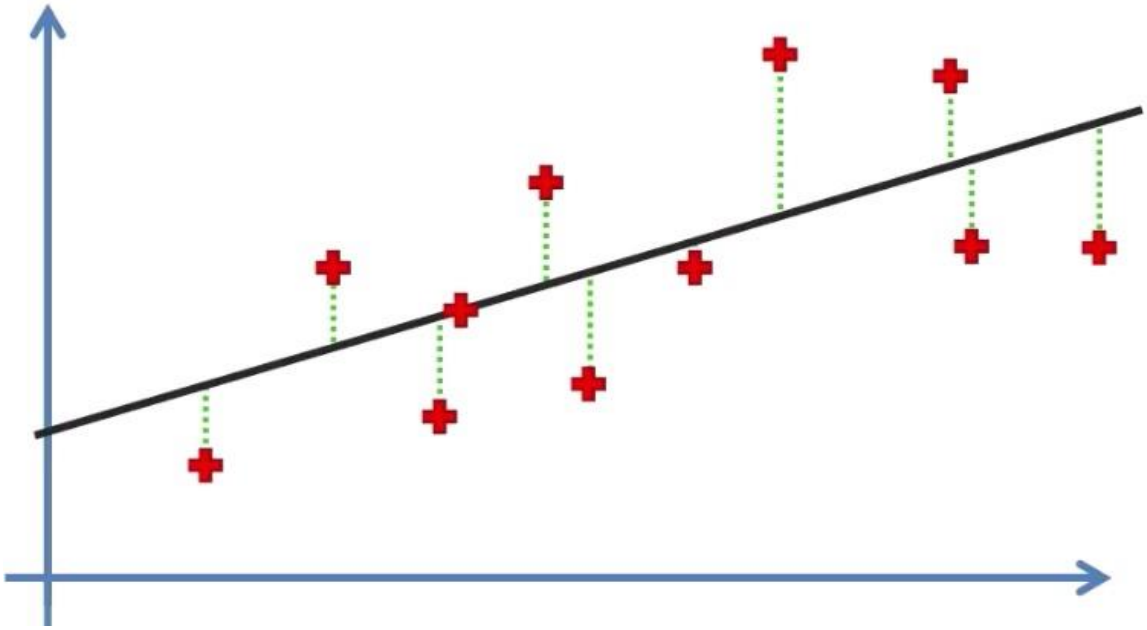


Рисунок 24

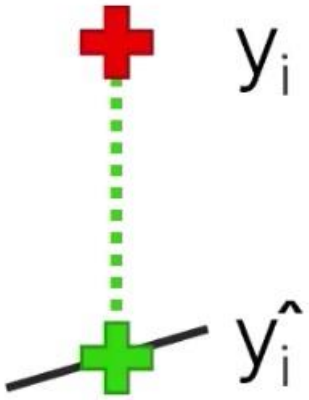



Рисунок 25, аркуш 29

де $y_i$ - істинне значення, $\hat{y_i}$ - зпрогнозоване значення заробітної платні. Формула похибок має наступний вигляд:

$$\sum_i (\hat{y_i} - y_i)^2 \to min \tag{2}$$

Узагальненням простої лінійної регресії є лінійна регресія:

$$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_n x_n \ , \tag{3}$$

де $y$ - залежна змінна,

$x_1, x_2, \ldots x_n$ - незалежні змінні,

$b_0, b_1, \ldots b_n$ -коефіцієнти.

Лінійна регресія має важливі припущення:

1. Лінійність;

2. Гомоскедастичність;
3. Багатомірна нормальність;
4. Незалежність похибок;
5. Відсутність мультиколінеарності.

На перший погляд, лінійна регресія може буди використана у роботі тільки з чисельними змінними, але це не так. Розглянемо приклад.

У таблиці 1 наведено дані про дохід 6 різних компаній за один рік, а також їх витрати за цей рік у доларах на R&D (науково-дослідні та дослідно-конструкторські роботи), адміністрування, маркетинг та штат у якому працює компанія.

Таблиця 1

| Дохід | R&D | Адміністрування | Маркетинг | Штат |
|---|---|---|---|---|
| 192 324 | 33 679 | 67 345 | 31 832 | New York |
| 156 767 | 29 845 | 48 434 | 29 212 | California |
| 189 347 | 18 364 | 51 459 | 35 694 | California |
| 134 670 | 22 654 | 50 432 | 21 942 | New York |
| 199 341 | 56 656 | 55 290 | 37 529 | Washington |

Як бачимо, штат не є чисельною змінною і в такому вигляді не може бути використаний у моделі. Перетворимо цю змінну на чисельну – для цього додамо фіктивні змінні до наших даних:

Таблиця 2

| Дохід | R&D | Адмін. | Марк. | New York | California |
|---|---|---|---|---|---|
| 192 324 | 33 679 | 67 345 | 31 832 | 1 | 0 |
| 156 767 | 29 845 | 48 434 | 29 212 | 0 | 1 |
| 189 347 | 18 364 | 51 459 | 35 694 | 0 | 1 |
| 134 670 | 22 654 | 50 432 | 21 942 | 1 | 0 |
| 199 341 | 56 656 | 55 290 | 37 529 | 0 | 0 |

У таблиці ми маємо три штати. Нашими фіктивними змінними будуть New York та California, а належність до штату позначимо через 1 і 0. Необхідно звернути увагу на те, що кількість фіктивних змінних має бути на одиницю менше від кількості штатів. Так, ми би могли отримати наступне:

Таблиця 3

| Дохід | R&D | Адмін. | Марк. | New York | California | Washington |
|---|---|---|---|---|---|---|
| 192 324 | 33 679 | 67 345 | 31 832 | 1 | 0 | 0 |
| 156 767 | 29 845 | 48 434 | 29 212 | 0 | 1 | 0 |
| 189 347 | 18 364 | 51 459 | 35 694 | 0 | 1 | 0 |
| 134 670 | 22 654 | 50 432 | 21 942 | 1 | 0 | 0 |
| 199 341 | 56 656 | 55 290 | 37 529 | 0 | 0 | 1 |

У такому випадку (таблиця 3) змінна Washington буде напряму корелювати зі змінними New York та California, що призведе до неправильної отриманої лінійної регресії. Нам завжди вистачить фіктивних змінних на одну менше: в нашому випадку належність до штату Washington буде отримана двома нулями у колонках New York і California (таблиця 2).

Також варто додати, що лінійна регресія добре працює для прогнозування неперервної величини, такої як дохід компанії, але не підходить для задач класифікації. [2]

**Логістична регресія** визначається функцією

$$\ln\left(\frac{p}{1-p}\right) = b_0 + b_1x_1 + b_2x_2 + \cdots + b_nx_n\ , \qquad (4)$$

де $p$ - залежна змінна,

$x_1, x_2, \ldots x_n$ - незалежні змінні,

$b_0,\ b_1, \ldots b_n$ -коефіцієнти.

У випадку якщо маємо одну незалежну змінну, логістична регресія буде мати наступний вигляд:

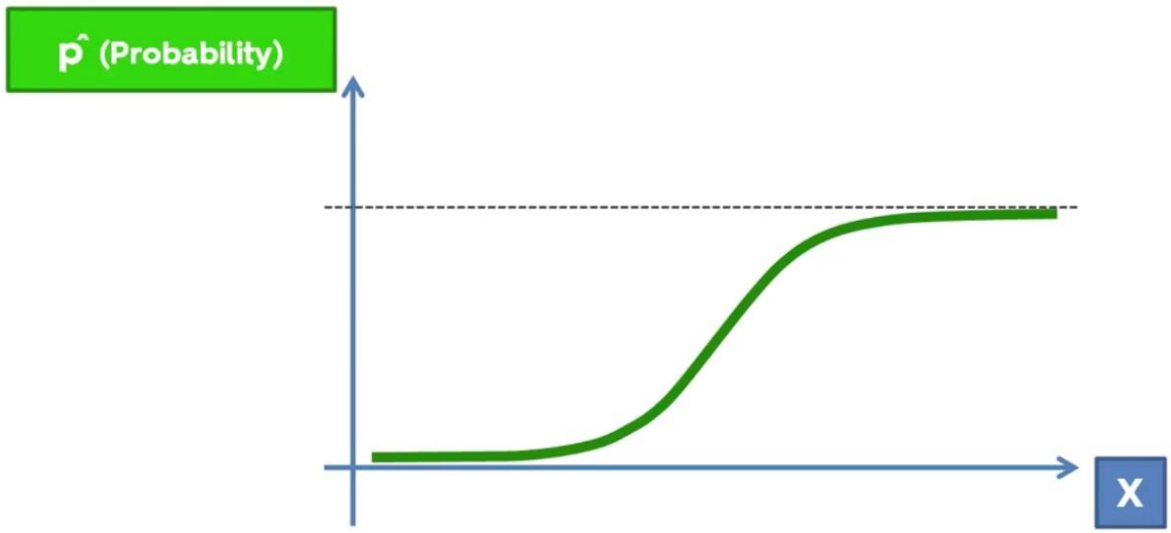



Рисунок 26

Вісь ординат – ймовірність, вісь абсцис – незалежна змінна $x$. Розглянемо приклад.

На рисунку 5 позначений набір даних клієнтів деякої компанії.

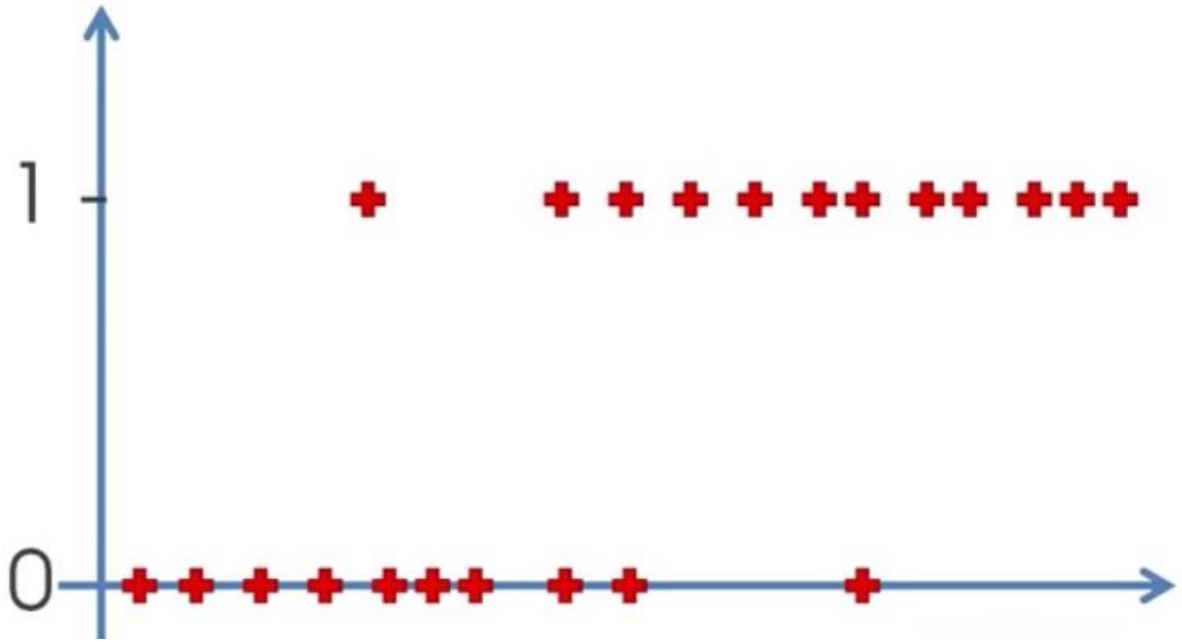



Рисунок 27, аркуш 33

На осі абсцис позначений вік клієнта, а на осі ординат – чи купив клієнт певний товар: 1 – клієнт купив товар, 0 – ні. При застосуванні логістичної регресії до набору даних ми отримаємо криву, яка буде відображати ймовірність покупки товару клієнтом. Щоб спрогнозувати покупку товару (так чи ні), позначимо поріг 0.5, значення більше якого будемо класифікувати як 1, а значення нижче – як 0:

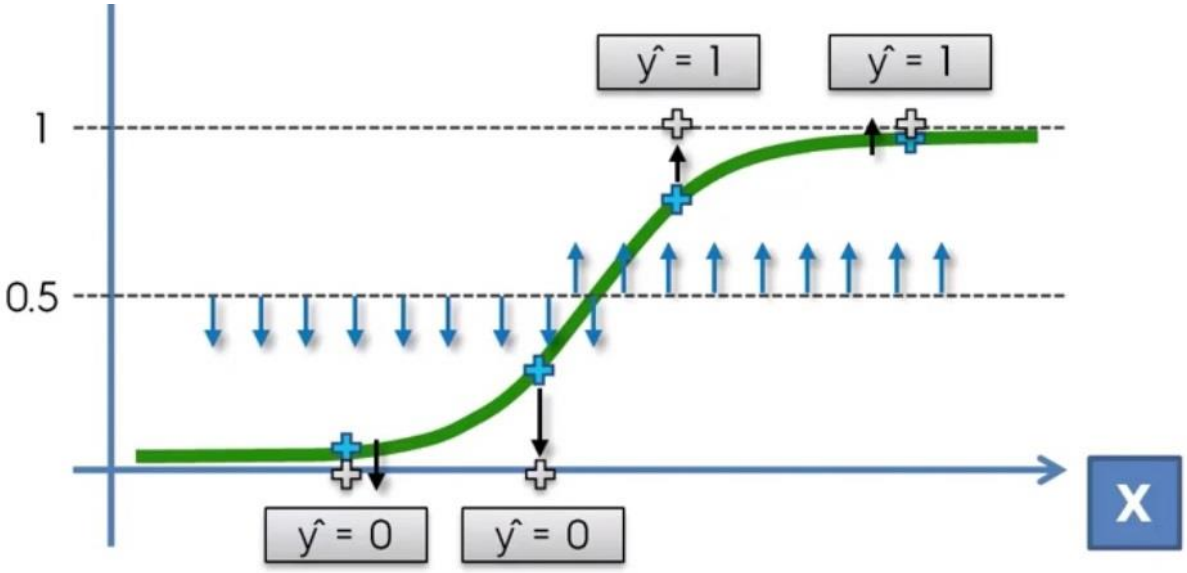



Рисунок 28

Зазвичай логістична регресія використовується для класифікації, я у розглянутому прикладі. [2]

## 2.2 Алгоритм k-NN

Алгоритм k-NN (k-Nearest Neighbors, k-найближчих сусідів), як і логістична регресія, може використовуватися для класифікації. Розглянемо приклад.

На рисунку 29 зображено деякий набір даних, точки якого розподілені на два класи, кожна точка має два значення – $x$ і $y$ , що відкладені по осям.

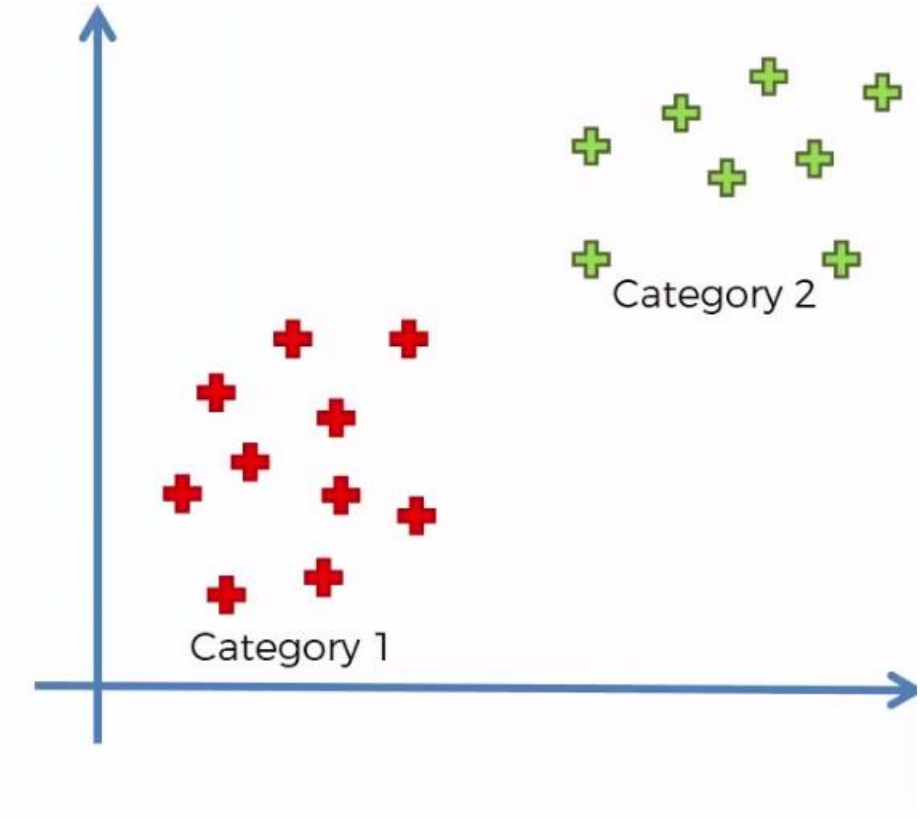



Рисунок 29

Нехай нам необхідно визначити, до якого з двох класів буде належати нова точка:

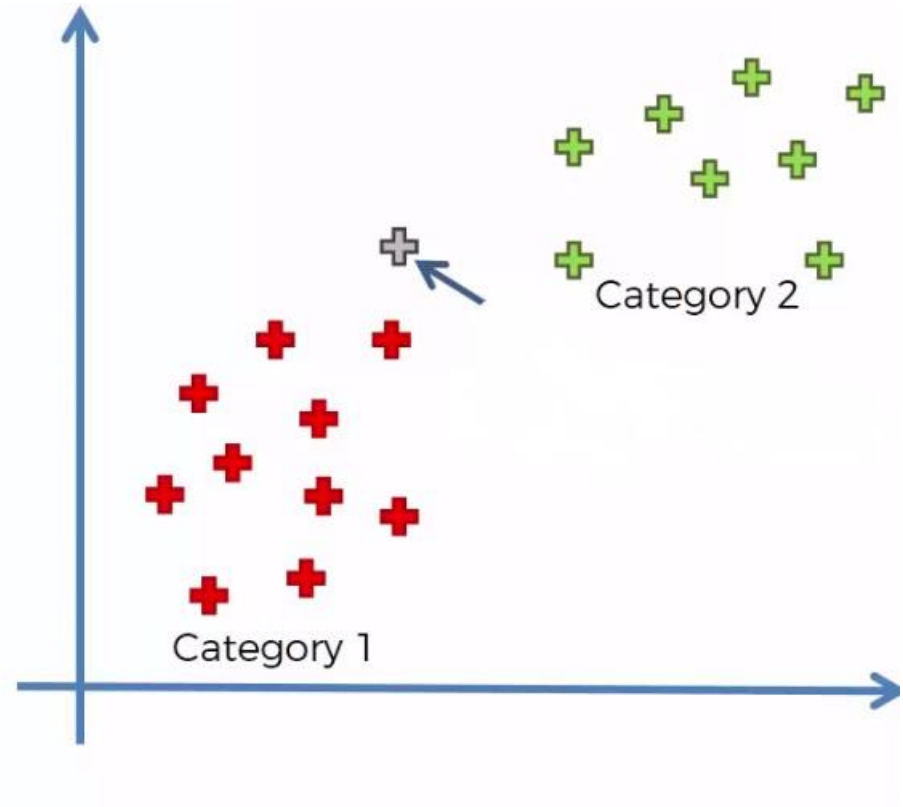



Рисунок 30, аркуш 35

Для цього скористаємося алгоритмом k-NN, робота якого полягає у виконанні над новою точкою наступних кроків:

1. Визначити число k.
2. Взяти k найближчих сусідів нової точки (найближчих з точки зору відстані до них).
3. Серед цих k сусідів порахувати, скільки з них належать до кожної категорії.
4. Призначити точку до тієї категорії, до якої належить більша кількість сусідів.

За відстань між двома точками можна взяти, наприклад, Евклідову відстань (для двомірного випадку):

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}\,, \tag{5}$$

де $(x_1, y_1)$ та $(x_2, y_2)$ відповідно координати першої та другою точки. [2]

## 2.3 Наївна модель Байєса

Наївна модель Байєса основана на відомій теоремі Байєса:

$$P(X|Y) = \frac{P(Y|X)\ P(X)}{P(Y)} \tag{6}$$

Для того, щоб робити передбачення, ми знаходимо ймовірність події, яка полягає у тому, що цільова ознака $t$ приймає певне значення $l$ за умови, що описові ознаки $q_1, \dots q_m$ також приймають певні значення. Узагальнена теорема Байєса для такої моделі буде виглядати наступним чином:

$$P(t = l \mid q_1, \dots q_m) = \frac{P(\ q_1, \dots q_m \mid t = l)\ P(t = l)}{P(\ q_1, \dots q_m)} \tag{7}$$

Отже, потрібно знайти три ймовірності:

1. $P(t = l)$ – **апріорна ймовірність** того, що цільова ознака $t$ приймає значення $l$;
2. $P(\ q_1, \dots q_m)$ – **сумісна ймовірність** того, що описові ознаки приймуть конкретний набір значень;
3. $P(\ q_1, \dots q_m \mid t = l)$ – **умовна ймовірність** того, що описові ознаки приймуть конкретний набір значень, за умови, що цільова ознака $t$ приймає значення $l$.

Перші дві ймовірності знайти відносно нескладно. $P(t = l)$ – це відносна частота, з якою цільова ознака приймає значення $l$ у наборі даних. $P(\ q_1, \dots\ q_m)$ знаходиться як відносна частота сумісної події, коли описові ознаки приймають певні значення. Крім того, цю ймовірність можна знайти як повну ймовірність (взявши суму по усім можливим значенням $k$ цільової ознаки $t$:

$$\sum_k P(\ q_1, \dots\ q_m \mid t = k)P(t = k) \tag{8}$$

Третю ймовірність $P(\ q_1, \dots\ q_m \mid t = l)$ можна знайти або безпосередньо з набору даних як відносну частоту події, або ж за допомогою ланцюгового правила. Ланцюгове правило дає можливість знайти ймовірність сумісної події як добуток умовних ймовірностей:

$$P(\ q_1, \dots\ q_m) = P(\ q_1) * P(\ q_2 \mid q_1) * \dots * P(\ q_m \mid q_{m-1}, \dots\ q_1) \tag{9}$$

Таким чином остаточно отримуємо:

$$P(\ q_1, \dots\ q_m \mid t = l) =$$

$$= P(\ q_1 \mid t = l) * P(\ q_2 \mid q_1,\ t = l) * \dots * P(\ q_m \mid q_{m-1}, \dots\ q_1,\ t = l) \tag{10}$$

Наївна модель Байєса зазвичай використовується для задач класифікації. [1]

## 2.4 Дерева рішень

Дерево прийняття рішень є дуже поширеною моделлю машинного навчання з відносно простою логікою побудови. Перейдемо одразу до прикладу.

Відома гра «Вгадай хто» — гра з двома гравцями, один з яких вибирає з колоди карту з зображенням персонажу, а інший намагається вгадати персонажа, задаючи питання. Відповіді на ці питання можуть бути лише «так» чи «ні». У таблиці 4 наведено 4 персонажі разом з деякими їх ознаками.

Таблиця 4

| Чоловік | Довге волосся | Окуляри | Ім’я |
|---|---|---|---|
| Так | Ні | Так | Браян |
| Так | Ні | Ні | Джон |
| Ні | Так | Ні | Афра |
| Ні | Ні | Ні | Аоифе |

Побудуємо послідовність питань, за допомогою яких можна вгадати одного з цих персонажів. Наприклад, можна було б побудувати таку послідовність:

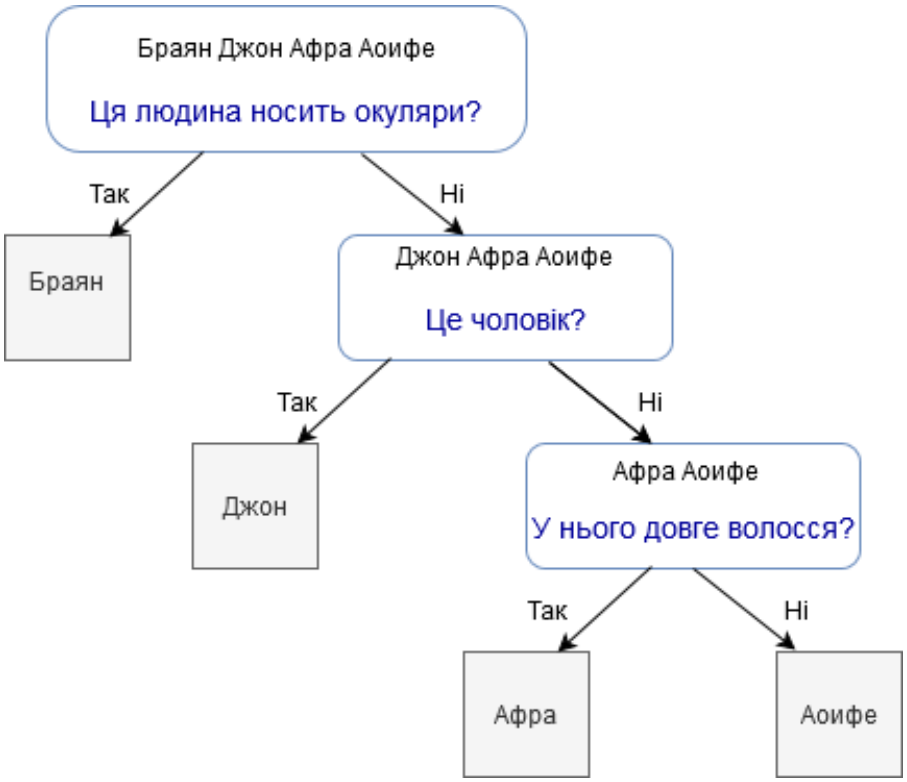



Рисунок 31, аркуш 39

Як бачимо, один шлях до персонажу має 1 питання, ще один має 2 питання, а інші – по 3 питання. Отже, якщо першим ставити питання «Ця людина носить окуляри?», середня кількість питань для відгадування персонажу буде дорівнювати

$$\frac{1+2+3+3}{4} = 2{,}25 \qquad (11)$$

Така ж середня кількість питань буде і в випадку, якщо першим будемо ставити питання «У цієї людини довге волосся?». А от якщо почати із запитання «Це чоловік?», отримаємо такий варіант:

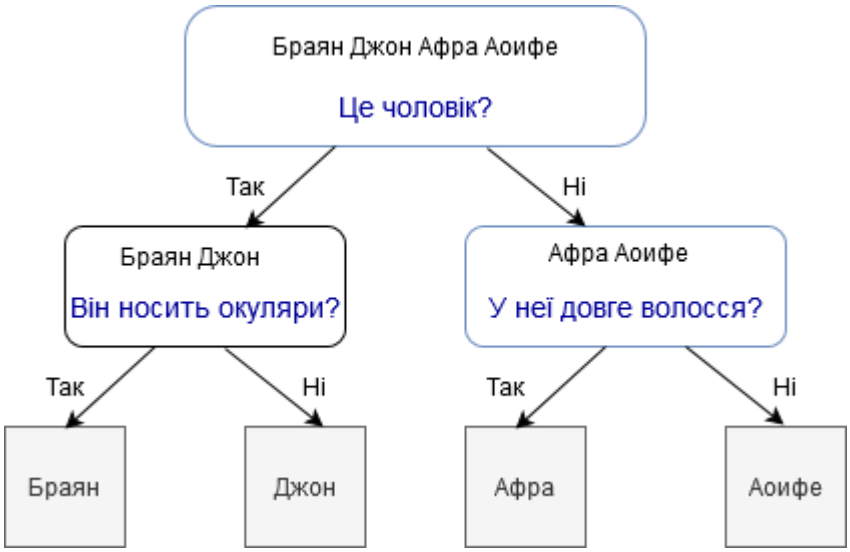



Рисунок 32, аркуш 40

В цьому випадку нам знадобиться всього 2 питання щоб вгадати будь якого персонажа, а тому і середня кількість буде дорівнювати 2. Така послідовність питань є найкращою для розглянутого набору даних. Цікаво, що відповідь на перше питання «Це чоловік?» дає нам більше інформації, аніж відповіді на інші два питання. Так відбувається не через саму відповідь «так» чи «ні», а через те, що це питання краще розбиває набір даних на дві множини. Таким чином, нам необхідно визначити, які описові ознаки є найбільш інформативними, щоб саме за допомогою них розбивати набір даних в першу чергу.

Дерево рішень складається із початкового вузла і внутрішніх вузлів, які зв’язані гілками. Кінцеві вузли називаються листами. Перший вузол являє собою початковий набір даних, який потім розбивається на частини з кожним новим вузлом. Листи являють собою частини даних, у яких є одна цільова ознака, яка і буде прогнозом для заданих описових ознак.

Кількісна величина, що показує наскільки добре ознака розбиває дані, вимірюється за допомогою моделі ентропії Шенона. Вона визначає кількісну міру неоднорідності елементів в множині:

$$H(t, D) = -\sum_{i \in l} \left[ P(t = i) * log_2\big(P(t = i)\big) \right], \quad (12)$$

де $P(t = i)$ – ймовірність того, що цільова ознака $t$ належить класу $i$, а $l$ – різні класи цільової ознаки у наборі даних $D$.

Мірою інформативності ознаки, яка використовується в деревах рішень, є приріст інформації. Формула (12) визначає ентропію для множини даних $D$ відносно цільової ознаки. Для того, щоб формально визначити приріст інформації, знадобиться ще одна формула:

$$rem(d, D) = \sum_{l \in \text{ рівні}(d)} \frac{|D_{d=l}|}{|D|} * H(t, D_{d=l}) \quad (13)$$

Формула (12) визначає ентропію, що залишилася після розбиття множини данних за допомогою конкретної ознаки $d$. Тепер можна визначити приріст інформації:

$$IG(d, D) = H(t, D) - rem(d, D). \quad (14)$$

Дерева рішень використовуються як для класифікації, так і для прогнозування неперервних велечин.

Окрім розглянутих моделей, також є відносно поширеними модель SVM та різноманітні моделі машинного навчання на основі нейромереж. [1]

## Висновки до розділу 2

В результаті виконаного аналізу моделей машинного навчання у першому розділі можна зробити висновок про велику різноманітність моделей і суттєву різницю у логіці їх побудови. Кожна модель має свої особливості, умови і межі використання. Вибір моделі і правильне її застосування є дуже важливим етапом в аналізі даних, бо це безпосередньо впливає на результат аналізу.

# РОЗДІЛ 3. СИСТЕМА ПРИЙНЯТТЯ РІШЕНЬ ДЛЯ ПРОГНОЗУ КРЕДИТНИХ ВИПАДКІВ НА ОСНОВІ МОДЕЛЕЙ МАШИННОГО НАВЧАННЯ

Після того, як таблиця даних готова до роботи, можна починати етап моделювання, на якому відбувається створення моделей машинного навчання на основі отриманих даних.

Вибір моделей. У розділі 2 було розглянуто декілька найбільш вживаних моделей машинного навчання. Будемо використовувати дерева рішень і наївну модель Байєса. По-перше, обидві ці моделі використовуються для класифікації і показують гарну точність прогнозів при правильній їх імплементації. По-друге, в основі цих моделей лежать кардинально різні математичні апарати: ентропія Шенона для дерев рішень і теорема Байєса для однойменної моделі. Буде цікаво порівняти, як два цих підходи впораються з поставленою задачею.

Варто додати, що для даного прогнозування окрім обраних алгоритмів можна використовувати такі моделі як логістична регресія або алгоритм k-NN.

Почнемо з моделі дерев рішень.

## 3.1 Дерева рішень для прогнозу кредитних випадків

Будемо використовувати модель Random Forest (Випадковий Ліс). Ця модель ґрунтується на алгоритмі побудови дерев рішень, основна його відмінність полягає у тому, що для прогнозування використовується не одне дерево, а множина дерев, кожне з яких будується на випадковій підмножині даних. Для класифікації нової точки, вона спочатку класифікується кожним деревом окремо, а остаточний клас визначається більшістю «голосів» дерев. Алгоритм побудови моделі виглядає наступним чином:

1. Вибрати випадкову кількість $k$ випадкових точок із тестової вибірки;

2. Побудувати дерево рішень на основі вибраних точок;

3. Обрати кількість дерев $n$ і повторити кроки 1 і 2 таку кількість разів;

4. Для отримання класу нової точки, спрогнозувати цю точку кожним із $n$ дерев і присвоїти їй той клас, що був обраний найбільше.

Побудуємо випадковий ліс із 10 деревами. Матриця похибок наведена на рис. 32:

| | 0 | 1 |
|---|---|---|
| 0 | 0.814662 | 0.607929 |
| 1 | 0.185338 | 0.392071 |

Рисунок 33 – Матриця похибок для 10 дерев

Як бачимо, точність прогнозування погашених кредитів (0 у матриці) є непоганою (81%), а от точність щодо дефолтних кредитів (1), що важливіше, є неприйнятно низькою – 39%. Це відбувається з тої причини, що дані є незбалансованими – менше чверті записів таблиці є дефолними кредитами. Спробуємо зрівняти кількість погашених та дефолтних кредитів у даних, зменшивши кількість погашених. Отримана матриця похибок:

| | 0 | 1 |
|---|---|---|
| 0 | 0.59714 | 0.372967 |
| 1 | 0.40286 | 0.627033 |

Рисунок 34 – Матриця похибок для збалансованої вибірки з 10 деревами

Хоч точність прогнозування погашених кредитів зменшилась, точність щодо дефолтних кредитів суттєво зросла до майже 63%. Спробуємо збільшити кількість дерев з 10 до 50:

| | 0 | 1 |
|---|---|---|
| 0 | 0.627457 | 0.367237 |
| 1 | 0.372543 | 0.632763 |

Рисунок 35, аркуш 45 – Матриця похибок для 50 дерев

Ще збільшимо кількість дерев:

| | 0 | 1 |
|---|---|---|
| 0 | 0.633173 | 0.368821 |
| 1 | 0.366827 | 0.631179 |

Рисунок 36 – Матриця похибок для 100 дерев

| | 0 | 1 |
|---|---|---|
| 0 | 0.644571 | 0.368535 |
| 1 | 0.355429 | 0.631465 |

Рисунок 37 – Матриця похибок для 300 дерев

Найкраща точність щодо прогнозування дефолтних кредитів, хоч і не суттєво краще за інші варіанти,  досягається у моделі з 50 деревами.

## 3.2 Наївна модель Байєса для прогнозу кредитних випадків

Застосуємо наївну модель Байєса до даних. Модель не потребує ніяких параметрів. Отримуємо матрицю похибок:

| | 0 | 1 |
|---|---|---|
| 0 | 0.844806 | 0.650699 |
| 1 | 0.155194 | 0.349301 |

Рисунок 38 – Матриця похибок моделі Байєса

Як і для моделі випадкового лісу, через незбалансованість даних точність прогнозування дефолту кредиту неприйнятно низька. Зрівняємо кількість погашених да дефолтних кредитів:

| | 0 | 1 |
|---|---|---|
| 0 | 0.638528 | 0.385054 |
| 1 | 0.361472 | 0.614946 |

Рисунок 39 – Матриця похибок моделі Байєса для збалансованої вибірки

Маємо точність прогнозування погашених кредитів майже 64%, а точність щодо дефолту – 61%.

Код python виконання моделей випадкового лісу і моделі Баєйса наведений у додатку А.

## 3.3 Аналіз результатів і ухвалення рішень

Найкращі результати для обох моделей наведені на рисунках нижче.

| | 0 | 1 |
|---|---|---|
| 0 | 0.627457 | 0.367237 |
| 1 | 0.372543 | 0.632763 |

Рисунок 40 – Матриця похибок дерев рішень

| | 0 | 1 |
|---|---|---|
| 0 | 0.638528 | 0.385054 |
| 1 | 0.361472 | 0.614946 |

Рисунок 41 – Матриця похибок наївної моделі Байєса

Побудуємо гістограму, що визначає наскільки великий внесок вносять поля у результат кредиту. Іншими словами, порівняємо поля за їх значимістю при ухваленні рішенння:

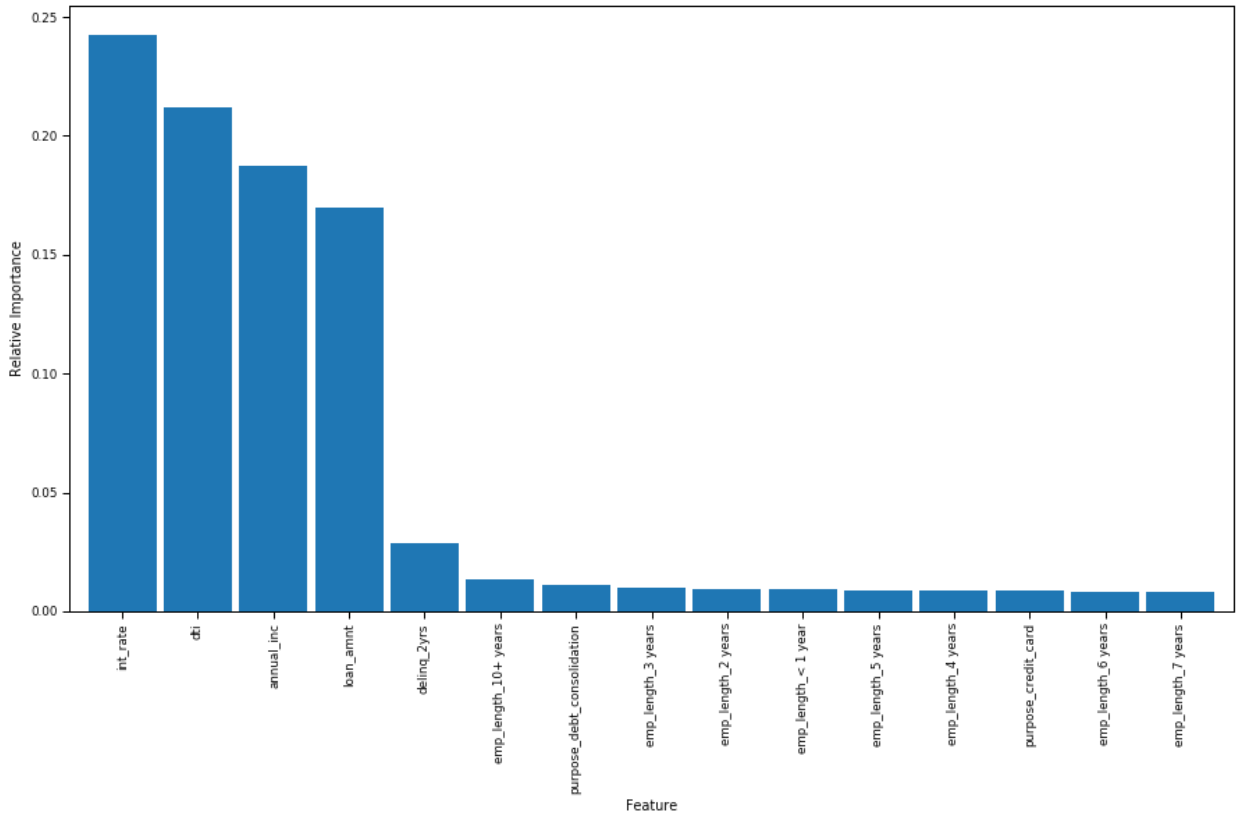



Рисунок 42, аркуш 47 – Гістограма значимості полів

Як бачимо, найбільша відносна значимість у полів int_rate (відсоткова ставка), dti (відношення боргу до доходу), annual_inc (річний дохід) і loan_amnt (розмір самого кредиту). Інші показники вносять відносно невиликий вклад (до 4%) у прийняття рішення.

Побудуємо ROC-криві моделей:

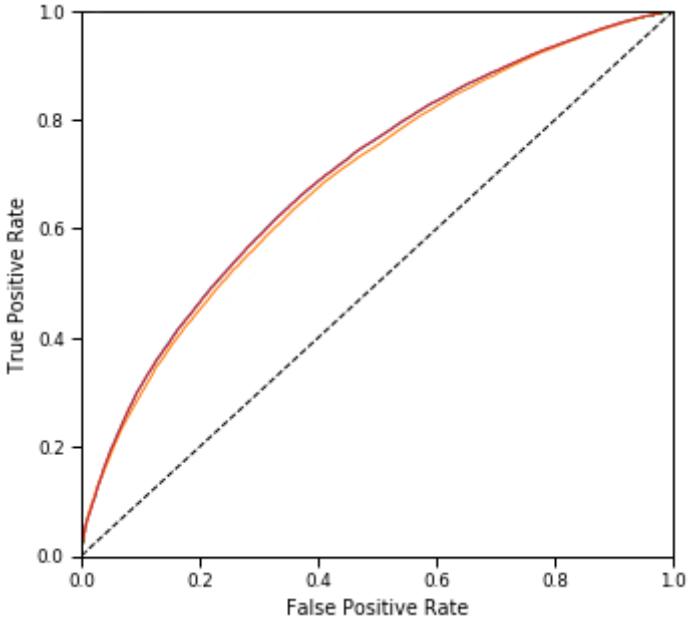



Рисунок 43, аркуш 48 – ROC-криві моделей (червоний – дерева рішень, помаранчевий – наївна модель Байєса)

Отримані криві майже співпадають, невелика перевага є моделі дерев рішень.

Оберемо модель дерев рішень для побудови системи прийняття рішень, адже ця модель показала дещо вищу ефективність прогнозування, порівняно з наївною моделлю Байєса. Блок-схема отриманої системи:


Рисунок 44, аркуш 49 – Блок-схема системи прийняття рішень

Висновки до розділу 3

Обрані моделі показали схожу точність прогнозування дефолту: 63% і 61% для дерев рішень і наївної моделі Байєса відповідно. Для покращення точності прогнозування можна розширити і переглянути простір описових ознак; збалансувати дані відносно кількості різних випадків, наприклад відносно розміру кредиту чи відсоткової ставки – прибрати перевагу одних значень над іншими; побудувати моделі на основі даних, зібраних за декілька останніх років, а не на основі усієї вибірки.

# РОЗДІЛ 4. ТЕХНІКО-ЕКОНОМІЧНЕ ОБҐРУНТУВАННЯ ТА ПИТАННЯ ОРГАНІЗАЦІЇ ВИРОБНИЦТВА

## 4.1 Постановка задачі проектування

Провести аналіз даних організації-кредитора, вибрати моделі машинного навчання для класифікації кредитів. Виконати порівняльний аналіз обраних моделей. Спроектувати і реалізувати систему прийняття рішень для прогнозування результату кредитів.

## 4.2 Обґрунтування функцій та параметрів програмного продукту

F1 – завантаження а) за допомогою файлу, б) введення вручну;

F2 – моделювання процесу та розрахунок прогнозу та попередня обробка даних а) реалізація моделі без визначення найкращих коефіцієнтів б) реалізація моделі з алгоритмом автоматичного визначення коефіцієнтів

F3 – вивід результатів прогнозу а) вивід за допомогою графіка, б) вивід за допомогою таблиці;

F4 – збереження результату роботи а) вивід лише на екран, б) збереження у файл

F5 – а) кешування даних, б) реалізація без кешування даних

Для розглянутих варіантів побудовано морфологічну карту (рис 4.1)

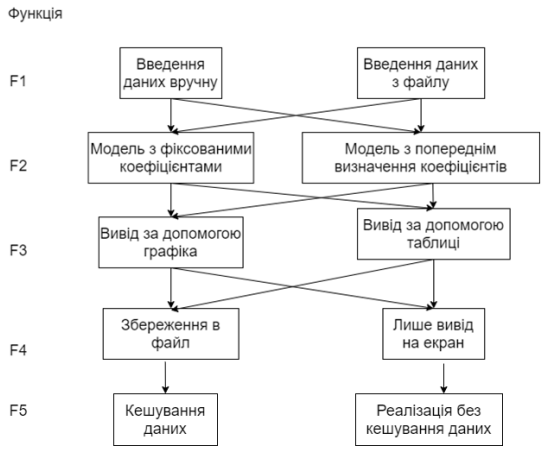



Рис. 4.1, аркуш 51 – Морфологічна карта

Для даної карти побудована позитивно-негативна матриця (табл. 4.1)

Таблиця 4.1 - Позитивно-негативна матриця

| Основна функція | Варіант реалізації | Переваги | Недоліки |
|---|---|---|---|
| F1 | а) | Не потрібно заздалегідь готувати дані | Незручно при великому об’ємі даних |
| | б) | Потрібно заздалегідь підготувати дані | Зручність при використанні даних |
| F2 | а) | Легкість в реалізації | Менша точність |
| | б) | Точність моделі зростає | Додаткова складність в реалізації |
| F3 | а) | Дозволяє швидко оцінити результати | Втрачається деталізація |
| | б) | Можливість легко деталізувати дані | Неочевидність загальної картини |
| F4 | а) | Можливість переглядати результати | Менша швидкість |
| | б) | Зручність при роботі та швидкодія | Немає можливості зафіксувати результати |
| F5 | а) | Результати попередніх розрахунків | Значний розмір програми |
| | б) | Незначний розмір програми | Програма не враховує попередні запуски |

Для характеристики прототипу програмного додатку використовуємо параметри X1 – X5. Визначаємо мінімальні, середні отримуванні та максимально допустимі значення (табл. 4.2)

Таблиця 4.2 – Система параметрів програмного продукту

| Найменування параметру | Позначення параметру | Значення параметру | | |
|---|---|---|---|---|
| | | Мінімальне | Середнє | Максимальне |
| Час розробки, людина*год | X1 | 200 | 325 | 450 |
| Час роботи алгоритму, мс | X2 | 200 | 5 100 | 10 000 |
| Потенційний розмір програми, МБ | X3 | 1 | 25 | 50 |
| Універсальність програми, частка одиниці | X4 | 1 | 0.75 | 0.25 |
| Зайнято б’ємо перативної пам’яті, МБ | X5 | 500 | 4250 | 8000 |

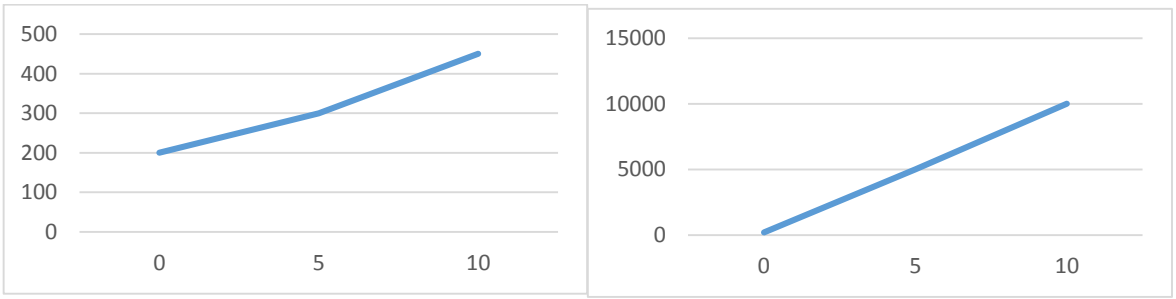



Рис. 4.2 – Значення параметрів час розробки та час роботи алгоритму

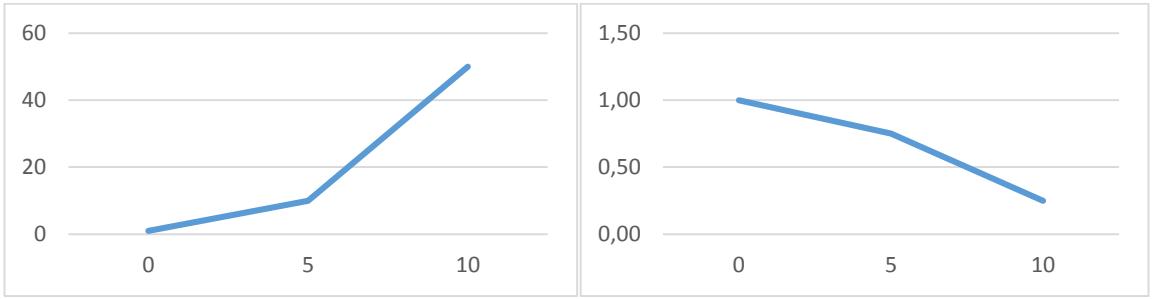



Рис. 4.3 – Значення параметрів потенційний розмір програми та універсальність програми

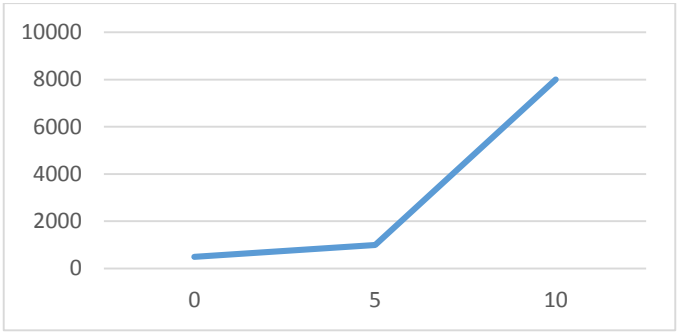



Рис. 4.4, аркуш 53 – Значення параметру зайнято б’ємо пеpативної пам’яті

Визначаємо ранги від 1 до 5, що визначаються експертною комісією:

Таблиця 4.3 – Результат оцінки параметрів

| Параметр | Ранг параметру по оцінці експерта | | | | | | | Сума рангів, $R_i$ | Відхилення $\Delta_i$ | Квадрат відхилення, $(\Delta_i)^2$ |
|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | | | |
| X1 | 2 | 1 | 1 | 2 | 1 | 2 | 2 | 11 | -10 | 100 |
| X2 | 1 | 2 | 3 | 1 | 3 | 3 | 1 | 14 | -7 | 49 |
| X3 | 3 | 4 | 4 | 4 | 5 | 5 | 3 | 28 | +7 | 49 |
| X4 | 5 | 3 | 2 | 3 | 2 | 1 | 4 | 20 | -1 | 1 |
| X5 | 4 | 5 | 5 | 5 | 4 | 4 | 5 | 32 | +11 | 121 |
| Разом | 15 | 15 | 15 | 15 | 15 | 15 | 15 | 105 | 0 | 320 |

Табл. 4.4 – Попарне зрівняння параметрів

| Параметри | Експерти | | | | | | | Кінцева оцінка | Числове значення |
|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | | |
| X1 та X2 | > | < | < | > | < | < | > | < | 0.5 |
| X1 та X3 | < | < | < | < | < | < | < | < | 0.5 |
| X1 та X4 | < | < | < | < | < | < | < | < | 0.5 |
| X1 та X5 | < | < | < | < | < | < | < | < | 0.5 |
| X2 та X3 | < | < | > | < | > | > | < | < | 0.5 |
| X2 та X4 | < | < | > | < | > | > | > | > | 1.5 |
| X2 та X5 | < | > | > | > | > | > | > | > | 1.5 |
| X3 та X4 | > | > | > | > | > | > | > | > | 1.5 |
| X3 та X5 | > | > | > | > | > | > | > | > | 1.5 |
| X4 та X5 | > | > | < | > | > | < | < | > | 1.5 |

Для експертних оцінок розрахуємо коефіцієнт конкординації

$$W = \frac{12S}{N^2(n^3-n)} = \frac{12*320}{7^2(5^3-5)} = 0.71 > W_k = 0.67 \text{ , де } S = \sum_{i=1}^{n} \Delta_i^2 = 320;$$

Отримане значення $W$більше за нормативне, то результати експертного оцінювання дозволяють продовжити дослідження.

Табл. 4.5 – Розрахунок вагомості параметрів

| Параметри Xi | Параметри Xj | | | | | Перший крок | | Другий крок | | Третій крок | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | X1 | X2 | X3 | X4 | X5 | $b_i$ | $К_{вi}$ | $b_i$ | $К_{вi}$ | $b_i$ | $К_{вi}$ |
| X1 | 1 | 0,5 | 0,5 | 1,5 | 1,5 | 5 | 0,2 | 22 | 0,194 | 100 | 0,2 |
| X2 | 1,5 | 1 | 0,5 | 1,5 | 1,5 | 6 | 0,24 | 27,5 | 0,239 | 124.75 | 0,24 |
| X3 | 1,5 | 1,5 | 1 | 1,5 | 1,5 | 7 | 0,28 | 34 | 0,296 | 155.50 | 0,28 |
| X4 | 0,5 | 0,5 | 0,5 | 1 | 1,5 | 4 | 0,16 | 17,5 | 0,152 | 80,25 | 0,16 |
| X5 | 0,5 | 0,5 | 0,5 | 0,5 | 1 | 3 | 0,12 | 14 | 0,122 | 64.50 | 0,12 |
| Загалом: | | | | | | 25 | 1 | 113 | 1 | 525 | 1,00 |

Враховуючи дані з порівнянь варіантів реалізацій функцій можна виключити з реалізацій функцій наступні варіанти: F1(б), F2(а), F4(б). Залишаються наступні варіанти:

1. F1(а)=>F2(а)=>F3(а)=>F4(б)=>F5(а)
2. F1(а)=>F2(а)=>F3(б)=>F4(б)=>F5(а)
3. F1(а)=>F2(а)=>F3(а)=>F4(б)=>F5(б)
4. F1(а)=>F2(а)=>F3(б)=>F4(б)=>F5(б)

Таблиця 5.6

| Основна функція | Варіант реалізації | Абсолютне значення параметру – | Бальна оцінка параметру – | Коефіцієнт вагомості параметру – | Коефіцієнт якості |
|---|---|---|---|---|---|
| F1 | а) | 400 | 8 | 0,2 | 1,6 |
| F2 | а) | 4000 | 4 | 0,24 | 0,96 |
| F3 | а) | 12 | 5,5 | 0,28 | 1,54 |
| | б) | 30 | 7 | 0,28 | 1,9 |
| F4 | а) | 0,92 | 8,9 | 0,16 | 1,42 |
| F5 | а) | 6000 | 8,5 | 0,12 | 1,02 |
| | б) | 1700 | 2,5 | 0,12 | 0.3 |

Обрахуємо коефіцієнти якості кожного з варіантів розробки:

$$K_{я1} = 1{,}6 + 0{,}96 + 1{,}54 + 1{,}42 + 1{,}02 = 6{,}54$$

$$K_{я2} = 1{,}6 + 0{,}96 + 1{,}9 + 1{,}42 + 1{,}02 = 6{,}9$$

$$K_{я3} = 1{,}6 + 0{,}96 + 1{,}54 + 1{,}42 + 0.3 = 5{,}82$$

$$K_{я4} = 1{,}6 + 0{,}96 + 1{,}9 + 1{,}42 + 0.3 = 6{,}18$$

Оскільки варіант 2 має найбільший коефіцієнт якості, то він є найкращим.

## 4.3 Економічний аналіз варіантів розробки

Для оцінки трудомісткості розробки спочатку проведемо розрахунок трудомісткості. Усі варіанти мають наступні основні завдання:

1) Введення даних з файлу
2) Модель з попереднім визначенням коефіцієнтів
4) Вивід результатів на екран

Також кожний з варіантів має два додатковихз авдання, які є реалізаціями розгалужених варіантів розробки незалежного модуля. Далі наведено варіанти додаткових завдань(два завдання, які мають номери 3 в реалізаціях та два завдання, які мають номери 5 в реалізаціях)

3.1) Вивід за допомогою графіку

3.2) Вивід за допомогою таблиці

5.1) Кешування даних

5.2) Реалізація без кешування даних

В варіанті 1 присутн інаступні додаткові завдання під номерами 4.1 та 5.1

В варіанті 2 присутні наступні додаткові завдання під номерами 4.2 та 5.1

В варіанті 3 присутні наступні додаткові завдання під номерами 3.1 та 5.2

В варіанті 4 присутні наступні додаткові завдання під номерами 3.2 та 5.2

За ступенем новизни до групи Б відноситься задання1, 2, 3.1, 3.2, 4, 5.1, 5.2

За складністю алгоритмів до групи 1 відносяться завдання 2, 5.1 до групи 2 відноситься завдання 1, 3 до групи 3 відноситься завдання 4.1, 4.2, 5.2.

Спираючись на норми розрахункового часу визначимо трудомісткість. Вона складає для першогозавдання $T_р$=5 людино-днів. Поправочний коефіцієнт складає $К_п$=0,91 (нормативно-довідкова інформація). Оскільки під час виконання даного завданняви користовуються новостворенні модулі, врахуємо це за допомогою коефіцієнта $К_{ст}$ = 0,8. Коефіцієнти $К_м$ і $К_{ст.п}$, які враховують відповідно програмування та розробку стандартного програмного забезпечення, для всіх семи завдань дорівнюють 1.

Повна трудомісткість 1 завдання (складність – 2, новизна – В):

$$T_1 = 5 * 0{,}91 * 0{,}8 = 3{,}64$$

Аналогічно для завдання 2 (складність – 2, новизна – В): –

$T_р$= 5; $К_п$ = 1,2; $К_{ст}$ = 0,8; $T_2 = 5 * 1{,}2 * 0{,}8 = 4{,}8$

Аналогічно для завдання 3(складність – 3, новизна – В):

$T_р$= 2; $К_п$ = 1; $К_{ст}$ = 0,6; $T_3 = 2 * 1 * 0{,}6 = 1{,}2$

Аналогічно для завдання 4.1(складність – 3, новизна – Г)

$T_р$= 2; $К_п$ = 0,56; $К_{ст}$ = 0,8; $T_4 = 2 * 0{,}56 * 0{,}8 = 0{,}9$

Аналогічно для завдання 4.2(складність – 3, новизна – Г)

$T_р$= 7; $К_п$ = 0,96; $К_{ст}$ = 0,6; $T_6 = 7 * 0{,}96 * 0{,}6 = 4{,}032$

Аналогічно для завдання 5.1(складність – 2, новизна – В):

$T_р$= 9; $К_п$ = 1,2; $К_{ст}$ = 0,8; $T_5 = 9 * 1{,}2 * 0{,}8 = 8{,}64$

Аналогічно для завдання 5.1(складність – 2, новизна – В):

$T_р$= 4; $К_п$ = 0,8; $К_{ст}$ = 0,8; $T_5 = 4 * 0{,}8 * 0{,}8 = 2{,}56$

Визначимоповнутрудомісткістьваріантів(людино-днів):

$$T_1 = 3{,}64 + 4{,}8 + 1{,}2 + 0{,}9 + 8{,}64 = 19{,}18$$

$$T_2 = 3{,}64 + 4{,}8 + 1{,}2 + 4{,}032 + 8{,}64 = 22{,}31$$

$$T_3 = 3{,}64 + 4{,}8 + 1{,}2 + 0{,}9 + 2{,}56 = 13{,}1$$

$$T_4 = 3{,}64 + 4{,}8 + 1{,}2 + 4{,}032 + 2{,}56 = 16{,}232$$

Найбільш трудомісткими завданнями є 1 і 2, найбільш трудомісткий варіант – 2.

Далі вважається, що робочий день складає 8 годин, в тижні п’ять робочих днів. В розробці бере участь один програміст з окладом 16 000 грн та тестувальник з окладом 10 000 грн .Визначимо середню заробітну плату за годину:

$$C_ч = \frac{16000 + 10000}{2 * 22 * 8} = 73{,}864$$

Тоді заробітна плата для кожного з варіантів реалізації(грн):

1) $C_{ЗП} = 73{,}864 * 8 * 19{,}18 = 11334{,}00$
2) $C_{ЗП} = 73{,}864 * 8 * 22{,}31 = 13183{,}00$
3) $C_{ЗП} = 73{,}864 * 8 * 13{,}1 = 7741{,}00$
4) $C_{ЗП} = 73{,}864 * 8 * 16{,}232 = 9592{,}00$

Відрахування на соціальнестрахування(22%)(грн):

1) $C_{ВІД} = 11334{,}00 * 0{,}22 = 2493$
2) $C_{ВІД} = 13183{,}00 * 0{,}22 = 2900$
3) $C_{ВІД} = 7741{,}00 * 0{,}22 = 1703$

4) $С_{ВІД} = 9592{,}00 * 0{,}22 = 2110$

Далі розрахуємо витрати на оплату однієї машино-години. Враховуючи, що вона обслуговує одного спеціаліста з окладом 16 000 грн та одного з окладом 10 000 грн з коефіцієнтомзайнятості 0,6, то для двох машин отримаємо

$$С_г = 12 * 16000 * 0{,}6 + 12 * 10000 * 0{,}6 = 187200 \text{ грн}$$

Враховуючидодатковузаробітну плату

$$С_{ЗП} = 187200 * (1 + 0{,}4) = 262080 \text{ грн}$$

Відрахування на соціальнестрахування 22%

$$С_{ВІД} = 270720 * 0{,}22 = 57658 \text{ грн}$$

Розрахуємо амортизаційні підрахунки (амортизація 25%, вартість ЕОМ 27 000 грн)

$$С_А = К_{ТМ} * К_А * Ц_{ПР} = 1{,}15 * 0{,}25 * 27\,000 = 7762{,}50 \text{ грн}$$

Розрахуємо витрати на ремонт та профілактику:

$$С_Р = К_{ТМ} * Ц_{ПР} * К_Р = 1{,}15 * 27000 * 0{,}05 = 1552{,}50 \text{ грн}$$

Розрахуємо ефективний годинний фонд часу ПК за рік

$$T_{ЕФ} = (365 - 142 - 16) * 8 * 0{,}8 = 1324{,}8 \text{ год}$$

Розрахуємо витрати на електроенергію

$$C_{ЕЛ} = 1324{,}8 * 0{,}6 * 0{,}78 * 1 * 3 = 1860{,}00 \text{ грн}$$

Накладні витрати рівні:

$$C_{Н} = 27000 * 0{,}67 = 18090 \text{ грн}$$

Отже експлуатаційні витрати(грн):

$$C_{ЕКС} = 262080 + 57658 + 7762{,}50 + 1552{,}50 + 1860{,}00 + 18090 = 349025{,}00$$

Тоді собівартість однієї машино-години ЕОМ дорівнюватиме:

$$C_{М-Г} = \frac{349025}{1324{,}8} = 263{,}45 \text{грн/год}$$

Враховуючи, що всі роботиведуться на ЕОМ, витрати на оплату машинного часу:

1) $С_м = 263,45 * 8 * 19,18 = 40332,30$
2) $С_м = 263,45 * 8 * 22,31 = 46913,50$
3) $С_м = 263,45 * 8 * 13,1 = 27546,68$
4) $С_м = 263,45 * 8 * 16,232 = 34132,65$

Накладні витрати відповідно

1) $С_Н = 40332,30 * 0,67 = 27022,64$
2) $С_Н = 46913,50 * 0,67 = 31432$
3) $С_Н = 27546,68 * 0,67 = 18456,28$
4) $С_Н = 34132,65 * 0,67 = 22870,00$

Розрахуємо повнувартість розробки за варіантами:

1) $С_{ПП} = 11334,00 + 2493 + 40332,30 + 27022,64 = 81181,94$
2) $С_{ПП} = 13183,00 + 2900,00 + 122236,17 + 81898,23 = 108281,14$
3) $С_{ПП} = 7741,00 + 1703,00 + 27546,68 + 18456,28 = 55446,96$
4) $С_{ПП} = 9592,00 + 2110,00 + 34132,65 + 22870,00 = 68704,65$

4.4 Вибір кращого варіанта ПП техніко-економічного рівня

$К_{я1} = 1,6 + 0,96 + 1,54 + 1,42 + 1,02 = 6,54$

$К_{я2} = 1,6 + 0,96 + 1,9 + 1,42 + 1,02 = 6,9$

$К_{я3} = 1,6 + 0,96 + 1,54 + 1,42 + 0.3 = 5,82$

$К_{я4} = 1,6 + 0,96 + 1,9 + 1,42 + 0.3 = 6,18$

$$К_{ТЕР1} = \frac{6,54}{81181,94} = 8,06 * 10^{-5}$$

$$К_{ТЕР2} = \frac{6,9}{108281,14} = 6,37 * 10^{-5}$$

$$К_{ТЕР3} = \frac{5,82}{55446,96} = 10,49 * 10^{-5}$$

$$К_{ТЕР4} = \frac{6,18}{68704,65} = 9 * 10^{-5}$$

Висновки до розділу 4

Отже, враховуючи всі дослідження, щоописанівище, можна сказати, що 3 варіант реалізації є найбільш оптимальним зі сторони якісно-економічної оцінки. Його коефіцієнт техніко-економічного рівня складає $10,49 * 10^{-5}$.

Розробка цього варіанту передбачає такіобов'язковіз авдання як:

1) Введення даних з файлу
2) Модель з попереднім визначенням коефіцієнтів
3) Вивід результатів на екран

Серед завдань між якими ставився вибір в даному варіанті реалізовані такі завдання: вивід за допомогою графіку; реалізація без кешування даних.

# ВИСНОВКИ ТА ПЕРСПЕКТИВИ ПОДАЛЬШИХ ДОСЛІДЖЕНЬ

На основі отриманих даних компанії «LendingClub» за 2007-2020 роки було побудовано систему прийняття рішень. Отримана система сформована на основі моделей машинного навчання і може бути використана компанією для ухвалення рішення про видачу кредиту фізичній особі в залежності від її даних. Однак модель машинного навчання, незалежно від точності її прогнозування, не може слугувати єдиним аргументом при прийнятті рішення. Річ у тім, що моделі показали максимальну точність лише 63%, що є доказом того, що неможливо взяти до уваги усі обставини і передумови – для двох випадків з однаковою передісторію один може виявитися дефолтом. У цьому випадку слід покладатися на досвід і професійну інтуїцію кредитора, що залишає місце для вдосконалення моделей. Подальший розвиток машинного навчання дозволяє проектувати алгоритми для аналізу якомога більшої кількості факторів, щоб робити точніші прогнози.

# СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Kelleher J.D.. Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies / Kelleher J.D., Namee B.M, D'Arcy A. – The MIT Press, 2015. – 624 p.
2. Eremenko K. Data Science A-Z: Real-Life Data Science Exercises Included [Електронний ресурс]. – 2020. – Режим доступу: https://www.udemy.com/course/datascience/, вільний.
3. Eremenko K. Machine Learning A-Z: Hands-On Python & R In Data Science / Eremenko K., de Ponteves H. [Електронний ресурс]. – 2020. – Режим доступу: https://www.udemy.com/course/machinelearning/, вільний.
4. Lending Club Loan Data [Електронний ресурс]. – 2020. – Режим доступу: https://www.kaggle.com/wendykan/lending-club-loan-data, вільний.

# ДОДАТОК А ЛІСТИНГ ПРОГРАМИ

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler, LabelEncoder, OneHotEncoder
from sklearn.compose import ColumnTransformer
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import confusion_matrix
from sklearn.naive_bayes import GaussianNB
from sklearn.metrics import roc_curve

PATH = 'C:\\Users\\loste\\Desktop\\data\\Loans_final.csv'

dataset = pd.read_csv(PATH)

y_default = dataset[dataset['loan_status'] == 1]
n_paid = dataset[dataset['loan_status'] == 0].sample(n=len(y_default))
dataset = y_default.append(n_paid)

X = dataset.iloc[:, 1:-1].values
```

```
y = dataset.iloc[:, 11].values



labelencoder_X = LabelEncoder()
X[:, 1] = labelencoder_X.fit_transform(X[:, 1])
X[:, 3] = labelencoder_X.fit_transform(X[:, 3])

ct = ColumnTransformer(
    [('one_hot_encoder', OneHotEncoder(categories='auto'), [4, 6, 7])],
    remainder='passthrough')
X = ct.fit_transform(X)

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.25)

sc = StandardScaler()
X_train = sc.fit_transform(X_train)
X_test = sc.transform(X_test)

# classifier = RandomForestClassifier(n_estimators = 300, criterion = 'entropy')
# classifier.fit(X_train, y_train)
```

```
classifier = GaussianNB()

classifier.fit(X_train, y_train)


y_pred = classifier.predict(X_test)


cm = confusion_matrix(y_test, y_pred, normalize='pred')



use [LoanData]

go


if object_id('Loans_bin') is not null

drop table [Loans_bin]


create table [Loans_bin]

(

        [row_number]            int identity(1,1),

        [loan_amnt]             float,

        [term]                  varchar(10),

        [int_rate]              float,

        [installment]           float,

        [grade]                         varchar(2),
```

```
	[sub_grade]		varchar(2),
	[emp_length]		varchar(10),
	[home_ownership]varchar(10),
	[annual_inc]		float,
	[verification_status] varchar(20),
	[purpose]		varchar(20),
	[dti]			float,
	[delinq_2yrs]		int,
	[loan_status_bin]	int

)

truncate table [Loans_bin]

insert into [Loans_bin]
(
	[loan_amnt],
	[term],
	[int_rate],
	[installment],
	[grade],
	[sub_grade],
```

```
        [emp_length],
        [home_ownership],
        [annual_inc],
        [verification_status],
        [purpose],
        [dti],
        [delinq_2yrs],
        [loan_status_bin]
)
select
        [loan_amnt],
        [term],
        [int_rate],
        [installment],
        [grade],
        [sub_grade],
        [emp_length],
        [home_ownership],
        [annual_inc],
        [verification_status],
        [purpose],
        [dti],
```

```
        [delinq_2yrs],
        case when [loan_status] = 'Fully Paid' then 0
                        when [loan_status] = 'Charged Off' then 1
                        when [loan_status] = 'Default' then 1
                        end as loan_status_bin

from [Loans_no_current]
```

# ДОДАТОК Б ДЕМОНСТАЦІЙНІ МАТЕРІАЛИ

## Система прийняття рішень в кредитуванні на основі методів машинного навчання

Роботу виконав
Печериця Віктор
гр. КА-64

Науковий керівник проф., д. т. н. В. Я. Данилов
Консультант проф., д. т. н. А. Б. Качинський

## Структура дослідження

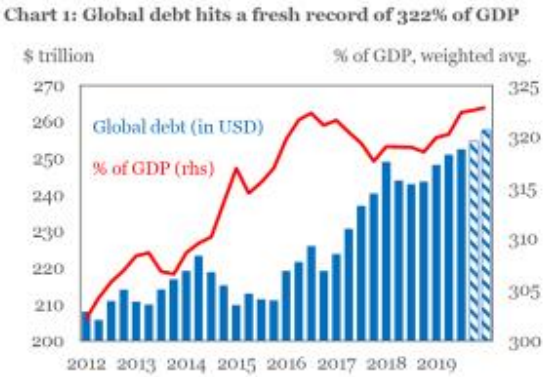


- Об’єкт дослідження – **кредитні випадки**, представлені статистичними даними
- Предмет дослідження – **процес аналітичного прогнозування**, що полягає у реалізації методів обробки статистичних даних і моделей машинного навчання на основі цих статистичних даних, а також аналізу результатів
- Мета роботи – побудова **системи прийняття рішень** для прогнозування результату кредиту на основі методів машинного навчання

## Актуальність

3

У 3-му кварталі 2019 р. **світовий борг** зріс до нового історичного максимуму у **$253 трлн**, що складає більше **320% світового ВВП**




Джерело: IFF (Інститут міжнародних фінансів США)

## Дані про кредитні випадки

4

LendingClub

**Сан-Франциско**, Каліфорнія

$10,8 млрд кредитів щороку у 2018 р.

Незабезпечений кредит
$1 000 - $40 000 на 3 роки

Станом на 2015 рік, **середній позичальник** має:

| | |
|---|---|
| credit score | 699 |
| дохід | $74 000 |
| кредит | $14 500 |
| ставка | 14,08% |

Джерело: сайт компанії lendingclub.com

## Аналітичне прогнозування

5

**Процес аналітичного прогнозування** складається з трьох основних частин:

1. Підготовка даних
2. Моделювання
3. Аналіз результатів

## Дані про кредитні випадки

6

### 1. Підготовка даних

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | | | | | |
| 1 | 1 | 2500 | 36 months | 13.56 | 84.92 | C | C1 | 10+ years | RENT | 55000 | Not Verified | debt_consolidation | 18.24 | 0 | Current |
| 2 | 2 | 30000 | 60 months | 18.94 | 777.23 | D | D2 | 10+ years | MORTGAGE | 90000 | Source Verified | debt_consolidation | 26.52 | 0 | Current |
| 3 | 3 | 5000 | 36 months | 17.97 | 180.69 | D | D1 | 6 years | MORTGAGE | 59280 | Source Verified | debt_consolidation | 10.51 | 0 | Current |
| 4 | 4 | 4000 | 36 months | 18.94 | 146.51 | D | D2 | 10+ years | MORTGAGE | 92000 | Source Verified | debt_consolidation | 16.74 | 0 | Current |
| 5 | 5 | 30000 | 60 months | 16.14 | 731.78 | C | C4 | 10+ years | MORTGAGE | 57250 | Not Verified | debt_consolidation | 26.35 | 0 | Current |
| 6 | 6 | 5550 | 36 months | 15.02 | 192.45 | C | C3 | 10+ years | MORTGAGE | 152500 | Not Verified | credit_card | 37.94 | 0 | Current |
| 7 | 7 | 2000 | 36 months | 17.97 | 72.26 | D | D1 | 4 years | RENT | 51000 | Source Verified | debt_consolidation | 2.4 | 0 | Current |
| 8 | 8 | 6000 | 36 months | 13.56 | 203.79 | C | C1 | 10+ years | RENT | 65000 | Source Verified | credit_card | 38.1 | 0 | Current |
| 9 | 9 | 8000 | 36 months | 17.97 | 180.69 | D | D1 | 10+ years | MORTGAGE | 63580 | Source Verified | debt_consolidation | 21.16 | 0 | Current |
| 10 | 10 | 8000 | 36 months | 14.47 | 206.44 | C | C2 | < 1 year | OWN | 300000 | Not Verified | debt_consolidation | 17.43 | 1 | Current |
| 11 | 11 | 3500 | 36 months | 22.35 | 211.09 | D | D5 | < 1 year | MORTGAGE | 90000 | Not Verified | credit_card | 15.84 | 1 | Current |
| 12 | 12 | 25000 | 60 months | 11.31 | 610.13 | B | B3 | 2 years | MORTGAGE | 70000 | Not Verified | credit_card | 22.01 | 0 | Current |
| 13 | 13 | 11200 | 36 months | 5.19 | 337.85 | A | A4 | 10+ years | MORTGAGE | 65000 | Not Verified | credit_card | 23.8 | 0 | Current |
| 14 | 14 | 6000 | 36 months | 17.97 | 214.9 | D | D1 | 4 years | MORTGAGE | 154000 | Source Verified | debt_consolidation | 28.78 | 2 | Current |
| 15 | 15 | 22000 | 60 months | 12.85 | 500.28 | B | B5 | 10+ years | MORTGAGE | 65000 | Source Verified | debt_consolidation | 11.19 | 0 | Current |
| 16 | 16 | 3600 | 36 months | 16.14 | 126.3 | C | C4 | 10+ years | MORTGAGE | 80000 | Verified | debt_consolidation | 13.63 | 0 | Current |
| 17 | 17 | 7000 | 36 months | 12.85 | 235.8 | B | B5 | 4 years | MORTGAGE | 102500 | Not Verified | house | 15.2 | 0 | Current |
| 18 | 18 | 25000 | 60 months | 16.91 | 620.11 | C | C5 | 10+ years | MORTGAGE | 23570 | Not Verified | debt_consolidation | 6.26 | 0 | Current |
| 19 | 19 | 16000 | 60 months | 20.89 | 431.67 | D | D4 | 4 years | MORTGAGE | 120000 | Not Verified | credit_card | 27.57 | 1 | Current |
| 20 | 20 | 10000 | 60 months | 14.47 | 305.67 | C | C2 | 10+ years | MORTGAGE | 75000 | Not Verified | debt_consolidation | 26.16 | 0 | Current |
| 21 | 21 | 10000 | 36 months | 13.56 | 339.65 | C | C1 | < 1 year | MORTGAGE | 65000 | Not Verified | credit_card | 10.62 | 0 | Current |
| 22 | 22 | 13000 | 36 months | 14.47 | 447.29 | C | C2 | 10+ years | MORTGAGE | 55000 | Verified | credit_card | 18.56 | 0 | Current |
| 23 | 23 | 9600 | 36 months | 23.4 | 373.62 | E | E1 | 9 years | RENT | 65000 | Not Verified | credit_card | 23.01 | 1 | Current |
| 24 | 24 | 2500 | 36 months | 20.89 | 131.67 | D | D4 | 10+ years | MORTGAGE | 40000 | Source Verified | car | 9.09 | 0 | Current |
| 25 | 25 | 16000 | 60 months | 26.31 | 481.99 | E | E4 | < 1 year | RENT | 33000 | Source Verified | credit_card | 33.62 | 0 | Current |
| 26 | 26 | 15000 | 60 months | 14.47 | 352.69 | C | C2 | n/a | MORTGAGE | 30000 | Source Verified | debt_consolidation | 41.6 | 0 | Current |
| 27 | 27 | 13000 | 36 months | 23.4 | 505.95 | E | E1 | 2 years | MORTGAGE | 99000 | Verified | other | 38.73 | 0 | Current |
| 28 | 28 | 23000 | 60 months | 20.89 | 620.01 | D | D4 | 5 years | RENT | 68107 | Source Verified | debt_consolidation | 0.52 | 0 | Current |
| 29 | 29 | 8000 | 36 months | 23.4 | 311.35 | E | E1 | 10+ years | OWN | 43000 | Source Verified | debt_consolidation | 33.24 | 0 | Current |
| 30 | 30 | 32075 | 60 months | 11.8 | 710.28 | B | B4 | 10+ years | MORTGAGE | 150000 | Not Verified | credit_card | 22.21 | 0 | Current |
| 31 | 31 | 12000 | 60 months | 13.56 | 276.49 | C | C1 | < 1 year | MORTGAGE | 40000 | Not Verified | debt_consolidation | 19.23 | 0 | Current |
| 32 | 32 | 10000 | 60 months | 19.82 | 264.5 | D | D3 | 10+ years | MORTGAGE | 80000 | Not Verified | debt_consolidation | 26.67 | 0 | Current |
| 33 | 33 | 16000 | 60 months | 17.97 | 406.04 | D | D1 | 5 years | MORTGAGE | 51000 | Not Verified | debt_consolidation | 21.91 | 0 | Current |

Таблиця має **2 260 000** записів

**Поля таблиці:**

1. Розмір кредиту
2. Срок кредиту
3. Відсоткова ставка
4. Щомісячний платіж
5. Клас

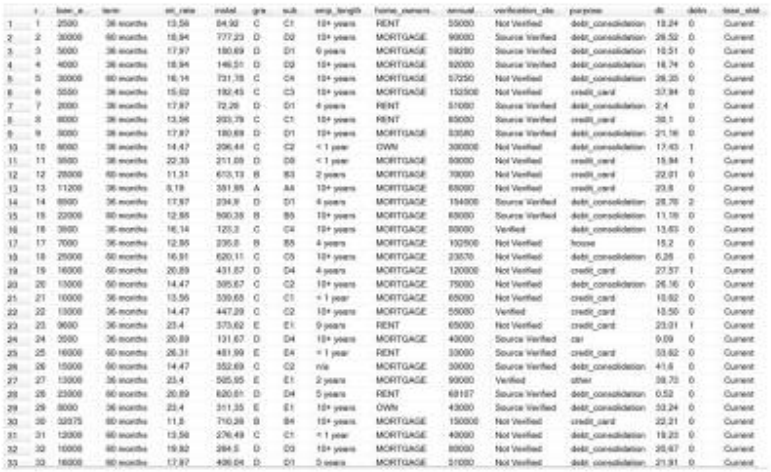
   (A, B, C, D, E, F, G)
6. Субклас

   (A1 – A5, B1 – B5, … G1 – G5)

## Дані про кредитні випадки

7

### 1. Підготовка даних



| | i | loan_a | term | int_rate | instal | grа | sub | emp_length | home_owners | annual | verification_sta | purpose | dti | delin | loan_stat |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 2500 | 36 months | 13,56 | 84,92 | C | C1 | 10+ years | RENT | 55000 | Not Verified | debt_consolidation | 18,24 | 0 | Current |
| 2 | 2 | 30000 | 60 months | 18,94 | 777,23 | D | D2 | 10+ years | MORTGAGE | 90000 | Source Verified | debt_consolidation | 26,52 | 0 | Current |
| 3 | 3 | 5000 | 36 months | 17,97 | 180,69 | D | D1 | 6 years | MORTGAGE | 59280 | Source Verified | debt_consolidation | 10,51 | 0 | Current |
| 4 | 4 | 4000 | 36 months | 18,94 | 146,51 | D | D2 | 10+ years | MORTGAGE | 92000 | Source Verified | debt_consolidation | 16,74 | 0 | Current |
| 5 | 5 | 30000 | 60 months | 16,14 | 731,78 | C | C4 | 10+ years | MORTGAGE | 57250 | Not Verified | debt_consolidation | 26,35 | 0 | Current |
| 6 | 6 | 5550 | 36 months | 15,02 | 192,45 | C | C3 | 10+ years | MORTGAGE | 152500 | Not Verified | credit_card | 37,94 | 0 | Current |
| 7 | 7 | 2000 | 36 months | 17,97 | 72,26 | D | D1 | 4 years | RENT | 51000 | Source Verified | debt_consolidation | 2,4 | 0 | Current |
| 8 | 8 | 8000 | 36 months | 13,56 | 203,79 | C | C1 | 10+ years | RENT | 65000 | Source Verified | credit_card | 38,1 | 0 | Current |
| 9 | 9 | 9000 | 36 months | 17,97 | 180,69 | D | D1 | 10+ years | MORTGAGE | 63580 | Source Verified | debt_consolidation | 21,16 | 0 | Current |
| 10 | 10 | 8000 | 36 months | 14,47 | 206,44 | C | C2 | < 1 year | OWN | 300000 | Not Verified | debt_consolidation | 17,43 | 1 | Current |
| 11 | 11 | 3600 | 36 months | 22,35 | 211,09 | D | D5 | < 1 year | MORTGAGE | 90000 | Not Verified | credit_card | 15,94 | 1 | Current |
| 12 | 12 | 28000 | 60 months | 11,31 | 613,13 | B | B3 | 2 years | MORTGAGE | 70000 | Not Verified | credit_card | 22,01 | 0 | Current |
| 13 | 13 | 11200 | 36 months | 8,19 | 351,85 | A | A4 | 10+ years | MORTGAGE | 68000 | Not Verified | credit_card | 23,8 | 0 | Current |
| 14 | 14 | 8000 | 36 months | 17,97 | 284,9 | D | D1 | 4 years | MORTGAGE | 154000 | Source Verified | debt_consolidation | 28,76 | 2 | Current |
| 15 | 15 | 22000 | 60 months | 12,55 | 500,28 | B | B5 | 10+ years | MORTGAGE | 65000 | Source Verified | debt_consolidation | 11,19 | 0 | Current |
| 16 | 16 | 3600 | 36 months | 16,14 | 125,3 | C | C4 | 10+ years | MORTGAGE | 80000 | Verified | debt_consolidation | 13,63 | 0 | Current |
| 17 | 17 | 7000 | 36 months | 12,55 | 235,6 | B | B5 | 4 years | MORTGAGE | 102500 | Not Verified | house | 15,2 | 0 | Current |
| 18 | 18 | 25000 | 60 months | 16,91 | 620,11 | C | C5 | 10+ years | MORTGAGE | 23570 | Not Verified | debt_consolidation | 6,26 | 0 | Current |
| 19 | 19 | 16000 | 60 months | 20,89 | 431,87 | D | D4 | 4 years | MORTGAGE | 120000 | Not Verified | credit_card | 27,57 | 1 | Current |
| 20 | 20 | 10000 | 60 months | 14,47 | 395,67 | C | C2 | 10+ years | MORTGAGE | 75000 | Not Verified | debt_consolidation | 26,16 | 0 | Current |
| 21 | 21 | 10000 | 36 months | 13,56 | 339,63 | C | C1 | < 1 year | MORTGAGE | 65000 | Not Verified | credit_card | 10,62 | 0 | Current |
| 22 | 22 | 13000 | 36 months | 14,47 | 447,29 | C | C2 | 10+ years | MORTGAGE | 55000 | Verified | credit_card | 18,50 | 0 | Current |
| 23 | 23 | 9600 | 36 months | 23,4 | 373,62 | E | E1 | 9 years | RENT | 65000 | Not Verified | credit_card | 23,01 | 1 | Current |
| 24 | 24 | 3000 | 36 months | 20,89 | 121,67 | D | D4 | 10+ years | MORTGAGE | 40000 | Source Verified | car | 9,09 | 0 | Current |
| 25 | 25 | 16000 | 60 months | 26,31 | 481,99 | E | E4 | < 1 year | RENT | 33000 | Source Verified | credit_card | 33,62 | 0 | Current |
| 26 | 26 | 15000 | 60 months | 14,47 | 352,69 | C | C2 | n/a | MORTGAGE | 30000 | Source Verified | debt_consolidation | 41,6 | 0 | Current |
| 27 | 27 | 13000 | 36 months | 23,4 | 505,95 | E | E1 | 2 years | MORTGAGE | 90000 | Verified | other | 38,73 | 0 | Current |
| 28 | 28 | 23000 | 60 months | 20,89 | 620,01 | D | D4 | 5 years | RENT | 68107 | Source Verified | debt_consolidation | 0,53 | 0 | Current |
| 29 | 29 | 8000 | 36 months | 23,4 | 311,35 | E | E1 | 10+ years | OWN | 43000 | Source Verified | debt_consolidation | 33,24 | 0 | Current |
| 30 | 30 | 32075 | 60 months | 11,8 | 710,28 | B | B4 | 10+ years | MORTGAGE | 150000 | Not Verified | credit_card | 22,21 | 0 | Current |
| 31 | 31 | 12000 | 60 months | 13,56 | 276,49 | C | C1 | < 1 year | MORTGAGE | 40000 | Not Verified | debt_consolidation | 19,23 | 0 | Current |
| 32 | 32 | 10000 | 60 months | 19,82 | 264,5 | D | D3 | 10+ years | MORTGAGE | 80000 | Not Verified | debt_consolidation | 26,67 | 0 | Current |
| 33 | 33 | 16000 | 60 months | 17,97 | 406,04 | D | D1 | 0 years | MORTGAGE | 31000 | Not Verified | debt_consolidation | 21,91 | 0 | Current |

Таблиця має **2 260 000** записів

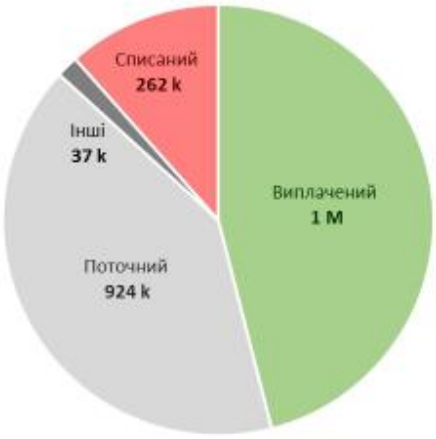
**Поля таблиці:**

7. Термін працевлаштування
8. Тип житла (іпотека, оренда, власність)
9. Дохід
10. Статус верифікації доходу
11. Ціль кредиту
12. Борг/дохід
13. Кількість кредитних правопорушень
14. Статус кредиту

## Дані про кредитні випадки

8

### 1. Підготовка даних




Розподіл статусів кредитів

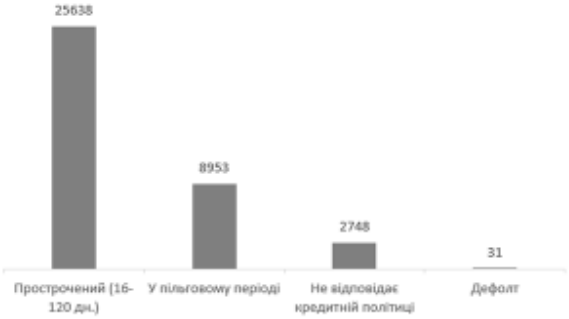
## Дані про кредитні випадки

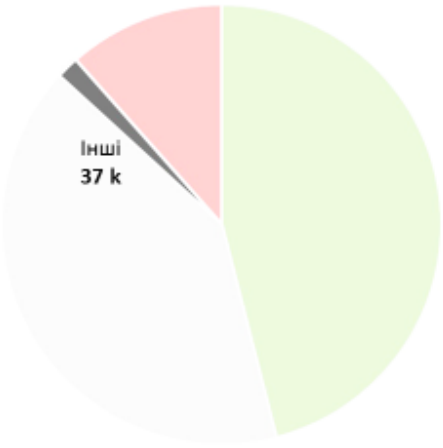
9

1. Підготовка даних

**Інші:**

1. Прострочений (16 – 120 дн.)
2. У пільговому періоді
3. Не відповідає кредитній політиці
4. Дефолт

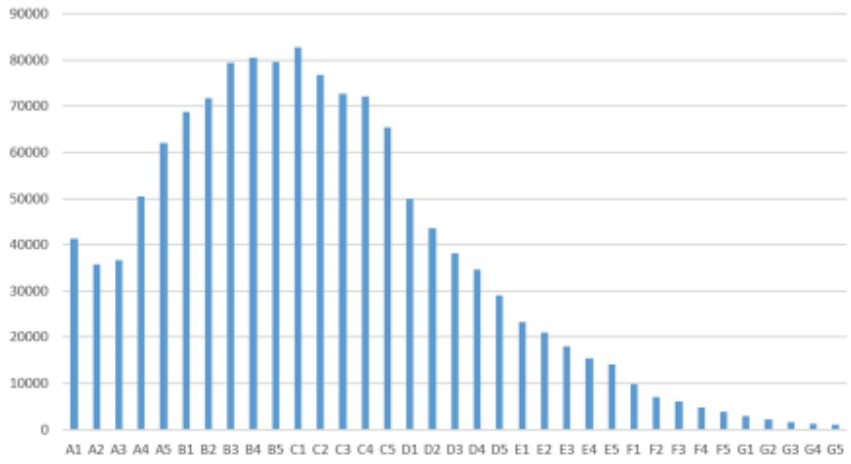



## Дані про кредитні випадки

10

1. Підготовка даних

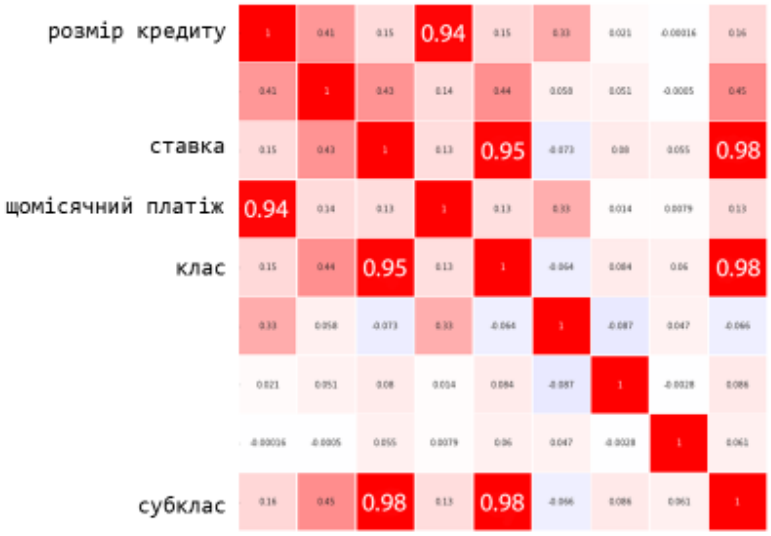



Розподіл субкласів кредитів

## Дані про кредитні випадки

11

### 1. Підготовка даних




**Кореляційна матриця**

## Моделювання

12

### Дерева рішень

Будемо використовувати модель **Random Forest**

1. Обрати кількість дерев $n$

$n$ разів

2. Вибрати випадкову кількість $k$ точок із вибірки
3. Побудувати **дерево рішень** на основі вибраних точок

4. Для отримання класу нової точки,
   спрогнозувати цю точку **кожним** із $n$ дерев
   і присвоїти їй той клас, що був обраний **найбільше**

Алгоритм побудови моделі

## Моделювання

13

### Наївна модель Байєса

$$P(t = l \mid q_1, \dots q_m) = \frac{P(q_1, \dots q_m \mid t = l)\, P(t = l)}{P(q_1, \dots q_m)}$$

$P(t = l)$ – ймовірність того, що цільова ознака $t$ приймає значення $l$

$P(q_1, \dots q_m)$ – сумісна ймовірність того, що описові ознаки $q_1, \dots q_m$ приймуть конкретний набір значень

$P(q_1, \dots q_m \mid t = l)$ – умовна ймовірність того, що описові ознаки приймуть конкретний набір значень, за умови що цільова ознака $t$ приймає значення $l$

## Аналіз результатів

14

Random Forest (10 дерев)

| | 0 | 1 |
|---|---|---|
| 0 | 0.814662 | 0.607929 |
| 1 | 0.185338 | 0.392071 |

Наївна модель Байєса

| | 0 | 1 |
|---|---|---|
| 0 | 0.844806 | 0.650699 |
| 1 | 0.155194 | 0.349301 |

Матриці похибок для **незбалансованої** вибірки

## Аналіз результатів

15

Random Forest

| | 0 | 1 |
|---|---|---|
| 0 | 0.59714 | 0.372967 |
| 1 | 0.40286 | 0.627033 |

10 дерев

| | 0 | 1 |
|---|---|---|
| 0 | 0.627457 | 0.367237 |
| 1 | 0.372543 | 0.632763 |

50 дерев

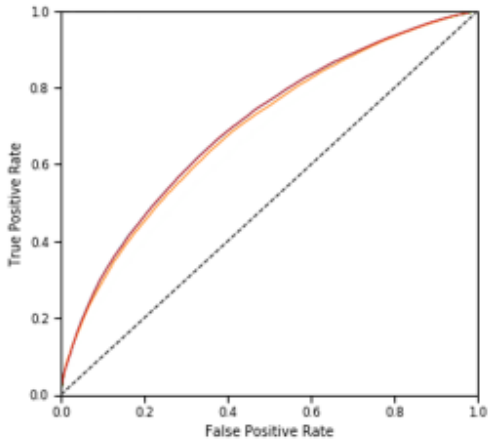
Наївна модель Байеса

| | 0 | 1 |
|---|---|---|
| 0 | 0.638528 | 0.385054 |
| 1 | 0.361472 | 0.614946 |

Матриці похибок для **збалансованої** вибірки

## Аналіз результатів

16




ROC-криві моделей

## Система прийняття рішень




## Дякую за увагу!

Роботу виконав
Печериця Віктор
гр. КА-64

Науковий керівник проф., д. т. н. В. Я. Данилов
Консультант проф., д. т. н. А. Б. Качинський